JOINT MODELING OF LONGITUDINAL MEASUREMENTS AND SURVIVAL DATA WITH COMPETING RISKS: APPLICATION TO HIV/AIDS STUDY

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in the Collaborative Graduate Program in Biostatistics

University of Saskatchewan

Saskatoon, Canada

By

Prosanta Kumar Mondal

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Graduate Chair, Collaborative Biostatistics Program

Department of Community Health & Epidemiology

University of Saskatchewan

104 Clinic Place

Saskatoon SK S7N 2Z4

Canada

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor Dr. Hyun J. Lim for her intelligent supervising, consistent support, motivation, and encouragement. I greatly appreciate her advice, time, and patience during the entire process of completing my PhD thesis.

I would like to thank my committee members: Dr. Punam Pahwa, Dr. Shahedul Khan, Dr. Chanchal Roy, and Dr. Suresh Tikoo for their expertise, encouragement, and taking effort in reading my thesis and providing me the valuable suggestion in completing my thesis.

My sincere thanks go to the School of Public Health, University of Saskatchewan and Western Regional Training Centre for Health Services Research (WRTC) for offering scholarship during my PhD program. I sincerely thank the Ontario HIV Treatment Network (OHTN) for allowing me to use the Ontario HIV Treatment Network Cohort Study (OCS) HIV data for my study.

Finally, I wish to thank my family: my two daughters, Modhurima and Shridula, my wife, Tumpa for their sacrifices and support during my study. I am grateful to my father Kalipado Mondal and mother Gouri Mondal for their unconditional love and encouragement. This thesis is dedicated to them.

ABSTRACT

Joint modeling of longitudinal measurements and survival data is a popular modeling technique in biomedical research (Wulfsohn and Tsiatis, 1997). Most of the studies in joint modeling consider only one failure type for the time-to-event outcome and an assumption of independent censoring. Some literature extends the methodology to allow for the multiple failures (also regarded as competing risks event) that frequently occur in clinical studies. However, only the Cox or other parametric cause-specific hazards (CSH) proportional survival submodels were used in those studies (Cox, 1972).

In this thesis, I study shared random effects joint models that consist of a linear mixed submodel for the longitudinal outcome, and Cox proportional CSH and proportional subdistribution hazards (SDH) submodels for the competing risks events (Fine and Gray, 1999; Laird and Ware, 1982; Rizopoulos, 2012). The longitudinal and the survival outcomes are linked together by latent random effects. To obtain estimates of the parameters, the joint likelihood of the longitudinal process and the survival process is used. The Expectation-Maximization (EM) algorithm was deployed to obtain maximum likelihood estimates of the parameters (Dempster, Laird, and Rubin, 1977).

I applied the methodology to a real HIV dataset that consisted of longitudinal biomarker CD4+ counts and cancer-related AIDS (cancer AIDS), and non-cancer AIDS as time-to-event outcomes. When cancer AIDS is the main event of interest, then non-cancer AIDS is a competing risk and vice versa. I compared results between joint models with the CSH and SDH submodels. For cancer AIDS, results in both the longitudinal and survival submodels varied

between the CSH-based and SDH-based joint models. However, for non-cancer AIDS, results were different in the longitudinal submodels but similar in the survival submodels. In my study population, proportions of individuals experiencing cancer AIDS and non-cancer AIDS were 2.7% and 15.0%, respectively. Thus, when non-cancer AIDS was the main event of interest, the proportion of competing event (cancer AIDS) was very low relative to non-cancer AIDS. Previous studies reported that if the proportion of individuals experiencing a competing risk is low, the CSH and SDH models may not provide different results. Hence, I conducted simulation studies to check the performance of the CSH and SDH models for different proportions of events and competing events. I observed that the results between CSH and SDH models are different if the proportion of individuals experiencing a competing risk is not much lower than the proportion experiencing the event of interest.

I also performed simulation study on the joint model to investigate how magnitudes of association parameter between longitudinal and survival outcomes influence the parameter estimates in separate Cox proportional hazards and linear mixed models. I observed that the bias of the estimate in separate Cox regression analysis increases as the magnitude of the association increases.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

ADC: Aids-Defining Conditions

ADI: Aids-Defining Illness

AIDS: Acquired Immunodeficiency Syndrome

ART: Antiretroviral Therapy

ARV: Antiretroviral

AZT: Azidothymidine

CDC: Centre for Disease Control and Prevention

cdf: cumulative distribution function

C.I.: Confidence Interval

CIF: Cumulative Incidence Function

CP: Coverage Probability

CSH: Cause-specific Hazards

EM: Expectation Maximization

FDA: Food and Drug Administration

GEE: Generalized Estimating Equations

HAART: Highly Active Antiretroviral Therapy

HIV: Human Immunodeficiency Virus

HR: Hazards Ratio

IDU: Injection Drug User

KS: Kaposi's Sarcoma

LME: Linear Mixed Effects

LRT: Likelihood Ratio Test

MLE: Maximum Likelihood Estimation

MSM: Men who have Sex with Men

NHL: Non-Hodgkin Lymphoma

NIAID: National Institute of Allergy and Infectious Diseases

OCS: Ontario Cohort Study

OI: Opportunistic Infection

pdf: probability density function

PH: Proportional Hazard

SDH: Subdistribution Hazards

VL: Viral Load

LIST OF KEY NOTATIONS

y_{ij}: Response variable for i^{th} individual at j^{th}measurement, $i = 1,2, \dots, n, j = 1,2, \dots, m_i$

t_{ij}: Time of j^{th}measurement for i^{th} individual or subject

n: Total number of individuals

m_i: Total number of measurements for i^{th} individual

$\boldsymbol{y}_i$: $(m_i \times 1)$ vector of responses or repeated measurements for i^{th} individual

V_i: Covariance matrix of y_i

$\boldsymbol{\beta}$: $(p \times 1)$ vector of fixed effects parameters

X_i: $(m_i \times p)$ known design matrix corresponding to fixed effects β

$\boldsymbol{\varepsilon}_i$: $(m_i \times 1)$ vector of measurement or sampling errors

R_i: Covariance matrix of ε_i

$\boldsymbol{b}_i$: $(q \times 1)$ vector of random effects for i^{th} individual

Z_i: $(m_i \times q)$ known design matrix corresponding to random effects b_i

G: Covariance matrix of b_i

$vech(G)$: Unique elements of covariance matrix G

T_i^*: Continuous random variable for event time/survival time/failure time for i^{th} individual, also indicates true event time

C_i: Censoring time for i^{th} subject

δ_i: Event indicator such that

$$\delta_i = \begin{cases} 1, & if\ the\ event\ was\ observed\ (T_i^* \leq C_i) \\ 0, & if\ response\ was\ censored\ (T_i^* > C_i) \end{cases}$$

T_i: Observed survival time for i^{th} subject

r: Number of failures/events

$t_{(g)}$: g^{th} ordered failure time, $g = 1,2, \dots, r$

n_g: Number of individuals who are alive just before time $t_{(g)}$

d_g: Number of individuals who fail at time $t_{(g)}$

$\boldsymbol{a}_i$: $(p \times 1)$ vector of covariates for i^{th} subject in Cox regression model

γ: Corresponding vector of regression coefficients

D: Censoring variable

e: Type of event, $e = 1,2, \dots, K$

CHAPTER 1

INTRODUCTION

1.1 Rationale

Joint modeling is an approach that is generally used to model simultaneously two response processes such as longitudinal (consists of repeated measurements) and survival (consists of time-to-event data) (Wulfsohn and Tsiatis, 1997). The overall objective of the joint analysis is to study the effect of covariates on the longitudinal outcome or survival outcome or on both. The joint modeling approach is quite often used in biomedical research. In clinical studies, repeated measurements on a response are generally obtained along with time-to-event outcome. One of the examples of joint modeling is found in Human Immunodeficiency Virus (HIV)/ Acquired Immunodeficiency Syndrome (AIDS) studies (Wulfsohn and Tsiatis, 1997). In these kinds of studies, biomarkers such as CD4+ count and viral load (VL) are measured at various occasions and time to AIDS or death is also recorded for each individual (Guo and Carlin, 2004; Wu, Liu, Yi, & Huang, 2012; Wulfsohn and Tsiatis, 1997). Another example of joint modeling is found in cancer studies (Pauler and Finkelstein, 2002; Proust-Lima & Taylor, 2009). For example, repeated prostate specific antigen measurements and time to disease recurrence are jointly modeled in prostate cancer studies (Pauler and Finkelstein, 2002).

Longitudinal measurements can be either continuous (e.g. Gaussian) or discrete (e.g. binary) (Salkind and Rasmussen, 2008). In this thesis, we focus on continuous repeated measurements. There are many well-known methods to analyze longitudinal and time-to-event data separately (Guo and Carlin, 2004). Linear Mixed Effects (LME) models based on maximum and restricted

maximum likelihood approach (Laird and Ware, 1982); Marginal and transitional models based on Generalized Estimating Equations (GEE) approach (Liang and Zeger, 1986) are the most commonly used methods for analyzing longitudinal data. We can have valid statistical inference from the longitudinal models based on likelihood methods or GEE approach if complete longitudinal data are available for all subjects or the missing data are missing at random (Song, 2013). However, in clinical studies, some participants are lost to follow-up because of disease progression or death or other reasons (Hunt and White, 1988). Thus, longitudinal measurements are not available from these dropout participants. In this case, these event-dependent dropouts can be directly related to what are being measured, and the missing values of longitudinal measurements caused by the dropouts of these participants can be informative and cannot be assumed as missing completely at random or missing at random (Little and Rubin, 2002; Rizopoulos, 2010).

Commonly used methods for analyzing survival data include the semiparametric Cox proportional hazards model (Cox, 1972) or a parametric model either with exponential or Weibull distributed (Weibull, 1951) baseline hazard function. However, one of the limitations of modeling longitudinal and survival outcomes separately is that the association between the two outcomes is generally ignored (Terrera, Piccinin, Johansson, Matthews, and Hofer, 2011). Separate analyses may not be appropriate when repeated measurements and time-to-event data are correlated (Guo and Carlin, 2004). Hence, an alternative approach of joint modeling of these two types of responses was proposed by several researchers (DeGruttola and Tu, 1994; Faucett and Thomas, 1996; Guo and Carlin, 2004; LaValley and DeGruttola, 1996; Pawitan and

Self, 1993; Taylor, Cumberland, & Sy, 1994; Tsiatis, DeGruttola, & Wulfsohn, 1995; Wulfsohn and Tsiatis, 1997).

Most joint models introduced in the literature so far have emphasized a single time-to-event or survival outcome (Guo and Carlin, 2004; Terrera et al., 2011; Wulfsohn and Tsiatis, 1997). However, we can have more than one survival outcome in our studies. In HIV studies, we define AIDS as a survival outcome and compare time-to-AIDS among different risk groups (Putter, Fiocco, & Geskus, 2007). However, a proportion of the HIV-infected patients could die before having AIDS. Here, death caused by other reasons before AIDS can be considered as a competing risk of AIDS (Geskus, 2016). In cancer studies, if our interest is on the death from cancer, then we can consider deaths because of other causes as competing risks of death from cancer (Bimali and He, 2015). As an alternative, if time to cancer relapse is the main event of interest, then death without relapse can be defined as a competing risk of relapse (Gooley, Leisenring, Crowley, and Storer, 1999).

A few researchers have proposed a joint model in competing risks scenarios in recent years (Chi and Ibrahim, 2006; Elashoff, G. Li, and N. Li, 2007 and 2008; Hu, G. Li, and N. Li, 2009; N. Li, Elashoff, and G. Li, 2009; Saville, Herring, and Koch, 2009; Williamson, Kolamunnage-Dona, Philipson, and Marson, 2008). In competing risks scenario, to study the effect of a covariate on a particular cause of failure, the traditional approach is to model the cause-specific hazards function, generally using a Cox cause-specific hazards model (Latouche, Boisson, Chevret, and Porcher, 2007; Prentice et al., 1978). Hence in the joint modeling of longitudinal and competing risks data, the cause-specific hazard submodel was mostly used (Elashoff et al., 2007, 2008; Saville et al., 2009; Williamson et al., 2008). However, several articles mentioned that a variable

can have different impact on the cause-specific hazard function, and on the cumulative incidence function for a specific type of failure (Fine and Gray, 1999; Gray, 1988; Latouche, Allignol, Beyersmann, Labopind, and Fine, 2013; Lunn and McNeil, 1995; Putter et al., 2007). If we are interested in the clinical prediction model, modeling the cumulative incidence could be preferable in some studies (Austin, Lee, and Fine, 2016; Dignam, Zhang, and Kocherginsky, 2012). Fine and Gray (1999) proposed a method for regression modeling with cumulative incidence functions (CIF) (M.-J. Zhang, X. Zhang, and Scheike, 2008). Their model is known as proportional subdistribution hazards model (Fine, 2001; Fine and Grey, 1999). In subdistribution hazards model, subjects who fail from a competing risk are not censored at the time of their failure but remain in the risk set (Lau, Cole, & Gange, 2009). This model compares the CIF of a particular cause of failure between groups of a covariate by modeling the subdistribution hazards (Latouche et al., 2007). Thus, in competing risks setting, joint modeling of the longitudinal and subdistribution hazards models has been proposed in recent years (Deslandes and Chevret, 2010).

However, there is very limited research done in the joint modeling of the longitudinal and subdistribution hazards models in competing risks situations. Deslandes and Chevret (2010) utilized joint modeling approach for both cause-specific and subdistribution hazards using data from an Intensive Care Unit (ICU). To the best of my knowledge, no joint modeling of longitudinal and subdistribution hazards models has been applied in the area of HIV/AIDS study. In this thesis, I applied joint modeling of longitudinal and subdistribution hazards models to the HIV/AIDS study. The study data of HIV-infected individuals was obtained from the Ontario HIV Treatment Network Cohort Study (OCS) (Rourke et al., 2013). The data consisted of

longitudinal outcome CD4+ counts and time-to-event outcomes of cancer AIDS or non-cancer AIDS. Kaposi's sarcoma and non-Hodgkin lymphoma were defined as AIDS-defining cancer or cancer AIDS (Ancelle-Park, 1993; Castro et al., 1992; Ebrahim, Abdullah, McKenna, and Hamers, 2004). All other AIDS-Defining Illnesses (ADI) were defined as non-cancer AIDS (Shiels et al., 2008, 2010). These two time-to-event outcomes were mutually exclusive. A person could be diagnosed either with cancer AIDS or with non-cancer AIDS. As experiencing one outcome/event might prevent a person from experiencing the other outcome, the two outcomes were considered as competing risks of each other (Putter et al., 2007). When cancer AIDS was the main event of interest then non-cancer AIDS was the competing risk, and vice versa (Shiels et al., 2008, 2010).

1.2 Study objectives and outline

The objectives of this study are as follows:

Objective 1: To compare the two joint modeling approaches based on (i) Cox cause-specific hazards and (ii) subdistribution hazards via their application to real HIV/AIDS data.

Objective 2: To examine the appropriateness of using cause-specific hazards and subdistribution hazards models based on the simulation study.

Objective 3: To investigate how magnitudes of association parameter between the longitudinal marker and time-to-event outcome influence parameter estimates obtained by conducting separate Cox regression analysis and linear mixed effects modeling from the simulation study.

This thesis is organized as follows: The literature review and statistical methods are provided in Chapter 2 and Chapter 3, respectively. I describe the analysis of longitudinal data, survival data, and competing risk in Section 3.2, Section 3.3, and Section 3.4, respectively. A joint modeling of longitudinal and survival data are presented in Section 3.5. I apply aforementioned methodologies to real HIV data in Chapter 4. The results from simulation studies are presented in Chapter 5. I presented discussion and conclusion in Chapter 6 and Chapter 7, respectively.

CHAPTER 2

LITERATURE REVIEW

2.1 Review of Joint Modeling Techniques

Joint modeling of longitudinal measurements and time-to-event data is a powerful and popular technique because the association between longitudinal and survival outcomes is considered in this method (Diaz, 2014; Ibrahim, Chu, and Chen, 2010; Sattar, Sinha, Argyropoulos, and Unruh, 2012). Several methods have been proposed to study the association between the two outcomes (Sweeting and Thompson, 2011). One of such methods is the extended Cox regression model, which allows us to model the time-dependent covariates to investigate their relationship to survival (Andersen and Gill, 1982). However, in the extended Cox regression analysis, we first need to determine whether time-dependent covariates are internal (also known as endogenous) or external (also known as exogenous) to the survival outcomes (Rizopoulos, 2012). Internal time-dependent covariates are subject-specific; the presence of the subject is required in the study for their measurements (Li and Ma, 2013). Therefore, we can observe an internal covariate up to the time point the subject is available (Balakrishnan and Rao, 2004; Dupuy and Mesbah, 2002). Internal covariates include, for example, the subject's white blood cell counts, CD4+ counts, blood pressure, and serum creatinine level (Li and Ma, 2013; Rizopoulos, 2012). On the other hand, to measure an external time-dependent covariate (e.g., air temperature, air pollution), we do not need to observe individual subjects (Li and Ma, 2013). The values of an external time-dependent covariate are not affected by the subject's failure process (Kalbfleisch and Prentice, 2002; Wang, 2004). Since

air pollution is not dependent on individual's asthma, air pollution is considered to be an external covariate to asthma (Rizopoulos, 2012). An external covariate is predictable, its value at any time t can be ascertained infinitesimally before t (Rizopoulos, 2012).

The extended/time-dependent Cox model is applicable for analyzing time-to-event data if the time-dependent covariates are external (Sattar et al., 2012). However, this model may not be appropriate if the time-dependent covariates are internal (Rizopoulos, 2012). The time-dependent Cox model assumes that time-dependent covariates can be predicted and they do not have any measurement error (Rizopoulos, 2012). In reality, the measurement error is not unusual as we cannot measure covariates perfectly (Bang et al., 2013). The time-dependent Cox model also assumes that the covariates change value at follow-up visits and are unchanged in between the visits (Rizopoulos, 2012). Because of such limitations of the time-dependent Cox model, joint modeling of longitudinal and survival outcomes was proposed and became popular in biomedical research (Sweeting and Thompson, 2011).

Joint modeling approaches were initiated mainly based on HIV/AIDS clinical studies, in particular, the joint modeling of longitudinal CD4+ counts and time-to-event data (Brown, Ibrahim, and DeGruttola, 2005; Chi and Ibrahim, 2006; DeGruttola and Tu, 1994; Faucett and Thomas, 1996; Faucett, Schenker, and Taylor, 2002; Ibrahim et al., 2010; LaValley and DeGruttola, 1996; Pawitan and Self, 1993; Taylor et al., 1994; Tsiatis et al., 1995; Wang and Taylor, 2001; Wulfsohn and Tsiatis, 1997). Several modifications of the joint model have been proposed in the literature (Faucett and Thomas, 1996; Wulfsohn and Tsiatis, 1997). Two major approaches have been proposed to perform joint modeling: a two-stage approach and a likelihood-based approach (Wu et al., 2012).

Two-stage joint modeling is the earlier approach (Hickey, Philipson, Jorgensen, and Kolamunnage-Dona, 2016). It is theoretically simpler than the likelihood based approach (Albert and Shih, 2010). Several two-stage approaches have been suggested in the joint modeling study (Dafni, 1993; Tsiatis et al., 1995; Wu et al., 2012). In early studies, two-stage procedures were used to determine whether CD4+ count could be considered a substitute marker for the time to AIDS or death (Dafni, 1993; Tsiatis et al., 1995; Murawska, 2013). Dafni (1993) and Tsiatis et al. (1995) developed a two-stage modeling approach considering survival as a function of a longitudinal covariate (Wulfsohn and Tsiatis, 1997). At the first stage, they modeled the longitudinal covariate by growth curve models with random intercept and slope (Laird and Ware, 1982; Wulfsohn and Tsiatis, 1997). In the second stage, they incorporated the modeled value obtained from the first stage as a time-dependent covariate in the Cox model (Wulfsohn and Tsiatis, 1997).

Many studies reported several drawbacks in the two-stage approach (Dafni and Tsiatis, 1998; Wulfsohn and Tsiatis, 1997). Wu et al. (2012) reviewed some of the two-stage methods (Dafni and Tsiatis, 1998; Ye, Lin, and Taylor, 2008) and concluded that biases might not be entirely eliminated by the two-stage methods.

Because of the drawbacks of the time-dependent Cox and two-stage approaches, Wulfsohn and Tsiatis (1997) proposed joint modeling using likelihood of the longitudinal and survival outcomes. This likelihood approach uses data efficiently from both the longitudinal and survival processes simultaneously (Yao, 2008). Hence, this approach makes proper statistical inferences about the impact of explanatory variables on the longitudinal and survival outcomes (McCrink,

Marshall, & Cairns, 2011). This likelihood approach also reduces biases in parameter estimation (Ibrahim et al., 2010).

Early studies on joint modeling (Wulfsohn and Tsiatis, 1997; Henderson, Diggle, and Dobson, 2000; Wang and Taylor, 2001) were reviewed by Tsiatis and Davidian (2004) and Yu, Law, Taylor, and Sandler (2004). Rizopoulos (2012) provided excellent discussions on joint modeling in his book (Sattar et al., 2012). Ibrahim et al. (2010) and Wu et al. (2012) reviewed joint models in a simpler way. Other researchers have also applied Bayesian techniques for joint modeling (Guo and Carlin, 2004; Huang, Dagne, and Wu, 2011; Ibrahim, Chen, and Sinha, 2004; Xu and Zeger, 2001).

Recent studies on joint modeling include Albert and Shih (2010), Ding and Wang (2008), Huang et al. (2011), Ibrahim et al. (2010), Lim, Mondal, and Skinner (2013), Rizopoulos (2011), Rizopoulos, Verbeke, and Lesaffre (2009), Sattar et al. (2012), Wu, Liu, and Hu (2010), Wu et al. (2012), and Ye et al. (2008). Ibrahim et al. (2010) discussed advantages of joint models by applying this methodology into metastatic breast cancer clinical trials. To examine the effect of treatment on overall survival, they performed analysis with and without using the repeated quality-of-life marker. Lim at al. (2013) considered a joint analysis with a longitudinal random effects submodel and a Cox or Weibull survival submodel to study the association between longitudinal CD4+ counts and the hazard of death among HIV-infected individuals. The hazard of death and the longitudinal process with an intercept and a slope was associated in their study. The parameter estimates were different in the joint model and separate models. Sattar et al. (2012) applied joint modeling approach on a hemodialysis study. They used serum albumin and all-cause mortality as longitudinal and time-to-event outcomes respectively in

their study. They reported that if there is a stronger relationship between longitudinal and survival outcomes, estimates in joint modeling may be more efficient compared to those from separate modeling (Sattar et al., 2012).

Joint modeling of longitudinal and competing risks data has been studied in recent years (Deslandes and Chevret, 2010; Elashoff et al., 2007 and 2008; Hu et al., 2009; Williamson et al., 2008). Elashoff et al. (2007) studied joint modeling of longitudinal measurements and competing risks data in a scleroderma lung disease clinical trial. Two time-to-event outcomes; treatment failure or death and disease-related dropout were considered in their study. They used linear mixed effects submodel with random intercept and slope for the longitudinal outcome and two-step mixture submodel for the survival outcomes. Similar covariates were considered in the longitudinal submodel and survival submodel. They used the EM algorithm to estimate the parameters of interest in their study (Dempster et al., 1977). Williamson et al. (2008) applied joint modeling approach to examine the effect of anti-epileptic drugs. They considered linear, piecewise mixed effects submodels for the longitudinal measurements and cause-specific hazards submodel for the multiple failure outcomes. Bayesian techniques are also applied in the joint analysis of longitudinal outcome and competing risks survival outcomes (Hu et al., 2009).

Deslandes and Chevret (2010) used joint models for both Cox cause-specific and subdistribution hazards in the ICU study. They considered sequential organ failure assessment score as the longitudinal outcome and time to discharge, and death as competing risks outcomes in their joint models. Unknown parameters of the model were estimated through the use of Markov chain Monte Carlo method of Gibbs sampling (Faucett and Thomas, 1996;

Metropolis and Ulam, 1949; S. Geman and D. Geman, 1984). They used the WinBUGS package (Spiegelhalter, Thomas, Best, and Lunn, 2003) for fitting their joint models.

Software: The **JM** package (Rizopoulos, 2010) in R software can be used to estimate most of the joint models for normal longitudinal measurements and time-to-event outcomes by a maximum likelihood approach (Wu et al., 2012). This package is suitable when our main interest is on the survival outcome, and we want to consider the impact of the longitudinal outcome on the survival outcome (Rizopoulos, 2010). Joint models of several types consisting of mixed-effects models and survival models with both parametric and semiparametric baseline hazards can also be fitted by the **JM** package (Sweeting and Thompson, 2011). The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is used to estimate parameters for Cox-type models (Sweeting and Thompson, 2011). A detailed description of the EM algorithm is provided in Chapter 4. Philipson, Sousa, and Diggle (2012) developed another package **joineR** in R for fitting joint models (Sweeting and Thompson, 2011). The **joineR** package fits a joint model where the focus is on both survival and longitudinal outcomes (Diaz, 2014).

2.2 Human Immunodeficiency Virus (HIV)

The Human Immunodeficiency Virus (HIV) is a virus that attacks the human immune system by reducing the number of T lymphocyte cells (HIV Monitoring, no date). T lymphocyte cells (CD4+ counts) help the body to fight infection (Edurant, 2016). CD4+ counts are measured based on the number of cells in a cubic millimeter (mm^3) of blood (Carter, 2016a). A healthy HIV-negative adult is assumed to have a CD4+ count of between 500 and 1500 cells/mm^3

(Carter, 2016a). The number of CD4+ counts is used to decide if someone has a weak immune system (WebMD, no date).

The initial stage of HIV infection is identified as primary HIV infection, or acute HIV infection (AIDSinfo, 2016). This stage continues as long as the body does not create antibodies against HIV (Healthline, 2015). At this initial stage, the immune system of the body is not prepared to fight the HIV (Carter, 2016b). As a result, large amounts of virus is produced in the body (AIDS.gov, 2015, August). The amount of virus in the blood is called the viral load (AIDS.gov, 2015, September). If someone gets HIV infection, the virus remains inside the body for entire life (Vyas, 2015). The transmission of HIV can happen from an infected person to another person through blood, semen, vaginal fluids, and breast milk (AIDS.gov, 2015a).

HIV attacks the CD4+ cells after entering the body, and more HIV is produced from infected CD4+ cells in the body (Nall, 2016). At the beginning, the body can produce adequate cells to maintain the number of CD4+ counts (Thaczuk, no date). However, as HIV continues to copy virus, the number of healthy CD4+ cells is decreased day by day (Knott, 2015). At one stage, the immune system of the infected person becomes weaker, and the body does not have sufficient capacity to fight infection (Thaczuk, no date). As a result, from HIV positive status, the person is progressed to Acquired Immune deficiency syndrome (AIDS) (AIDS.gov, 2015b).

AIDS is considered as the last stage of HIV disease (Cherney, 2014). If a person has one or more opportunistic infections (See 2.2.3), and a lower number of CD4+ counts (< 200), the person will be diagnosed with AIDS (HIV Monitoring, no date). After infection, times to develop AIDS are not equal for all infected-person (WHO, 2016). Without taking antiretroviral therapy (ART), almost all HIV-positive people will progress to AIDS (Vyas, 2015).

2.2.1 Epidemiology of HIV

Globally, about 35.3 (range: 32.2 to 38.8) million people had HIV in 2012 (UNAIDS, 2013). Global incidence has decreased from 3.4 million in 2001 to 2.3 million in 2012 (UNAIDS, 2013). Reductions in heterosexual transmission contribute this large reduction in global HIV incidence (Maartens, Celum, and Lewin, 2014).

HIV epidemics has two categories: concentrated and generalized (Green and Ruark, 2016). In "concentrated" epidemic, less than 1% HIV infection occurs from the general population, but more than 5% infection occurs from "high risk" groups (Denning and DiNenno, 2015). Men who have sex with men (MSM), sex workers, and injection drug users (IDUs) are considered to be "high risk" or vulnerable groups (Comiskey, Dempsey, Simic, and Baroš, 2013). The future course of "concentrated" epidemic depends on the magnitude of the vulnerable populations (UNAIDS, 2011). The epidemic of North America and the Caribbean is considered as "concentrated" (Volberding, Greene, Lange, Gallant, and Sewankambo, 2012). An epidemic is "generalized" when more than 1% HIV infection occurs among the general population (Denning and DiNenno, 2015). The generalized epidemic depends on the sexual behavior of the general population (Mishra et al., 2012).

2.2.2 HIV in Canada

According to a 2011 national HIV estimates, about 71,300 (range: 58,600 to 84,000) people had HIV in Canada (PHAC, 2014). About 50% of the infected-people was only MSM (Men who have Sex with Men), and either MSM or injection drug users (MSM-IDUs) (Challacombe, 2013). The number of new infections in each year did not change in the last decade (PHAC, 2014). In

2011, MSM and IDUs had 71 and 46 times higher incidence rates than the rates among their corresponding counterparts, respectively (PHAC, 2014). About twelve percent of all the new infections occur in the Aboriginal people (Challacombe, 2013). The modes of HIV transmission can vary among populations: Aboriginal people get infection mainly by contaminated needles (IDU), heterosexual contact is key among women (PHAC, 2014).

2.2.3 Opportunistic infections (OIs)

Opportunistic infections (OIs) are infections that happen due to the week immune system of a body (MedicineNet.com, 2016). Individuals with an advanced stage of HIV infection are susceptible to opportunistic infections (UNAIDS, 1998). The incidence of OIs has been reduced by the antiretroviral therapy (ART) (Shahapur and Bidri, 2014). All HIV-infected individuals should be aware of the common OIs to prevent them from these infections or to receive earlier treatment for OIs (CDC, 2016).

When an infected person gets certain OIs, the person is diagnosed with AIDS (AIDS.gov, 2010). AIDS is also diagnosed if the CD4+ count falls below 200 cells/mm^3 in persons with HIV (AIDS.gov, 2016). The CDC (2008) developed the following list of OIs that are considered as AIDS-Defining Conditions (ADC) or AIDS-Defining Illness (ADI):

- Bacterial infections, multiple or recurrent
- Candidiasis of bronchi, trachea, or lungs
- Candidiasis of esophagus
- Cervical cancer, invasive

- Coccidioidomycosis, disseminated or extrapulmonary
- Cryptococcosis, extrapulmonary
- Cryptosporidiosis, chronic intestinal (>1 month's duration)
- Cytomegalovirus disease (other than liver, spleen, or nodes), onset at age >1 month
- Cytomegalovirus retinitis (with loss of vision)
- Encephalopathy, HIV-related
- Herpes simplex: chronic ulcers (>1 month's duration) or bronchitis, pneumonitis, or esophagitis (onset at age >1 month)
- Histoplasmosis, disseminated or extrapulmonary
- Isosporiasis, chronic intestinal (>1 month's duration)
- Kaposi sarcoma
- Lymphoid interstitial pneumonia or pulmonary lymphoid hyperplasia complex
- Lymphoma, Burkitt (or equivalent term)
- Lymphoma, immunoblastic (or equivalent term)
- Lymphoma, primary, of brain
- Mycobacterium avium complex or Mycobacterium kansasii, disseminated or extrapulmonary
- Mycobacterium tuberculosis of any site, pulmonary, disseminated, or extrapulmonary
- Mycobacterium, other species or unidentified species, disseminated or extrapulmonary
- Pneumocystis jirovecii pneumonia
- Pneumonia, recurrent
- Progressive multifocal leukoencephalopathy

- Salmonella septicemia, recurrent
- Toxoplasmosis of brain, onset at age >1 month
- Wasting syndrome attributed to HIV

Regardless of an HIV-infected individual's CD4+ count, having a diagnosis with any of these OIs means the individual has progressed to AIDS (AIDS.gov, 2010). In this thesis, AIDS will be defined based on the above list of OIs by the CDC (2008).

2.2.4 HIV treatment

Antiretroviral therapy (ART) is used to treat HIV (AIDSinfo, 2016). ART prevents the HIV from replicating and from destroying the immune system of an infected person (AIDSinfo, 2016). Hence by using ART, the amount of viral load can be reduced in the body (AIDS.gov, 2015, September). In 1987, Food and Drug Administration (FDA) approved AZT (Zidovudine) as the first treatment for HIV (Aidsmap, no date). Since then the FDA has approved 31 Antiretroviral drugs (ARVs) to treat HIV-infected people (Elder, 2013). Different types of ARVs attack the virus in different ways (CATIE, no date). A combination of three or more ARV drugs is defined as Highly Active Antiretroviral Therapy (HAART) (WHO, no date). HAART is the most effective treatment of HIV/AIDS and has been widely available since 1997 (Touloumi et al., 2004).

Till now, no medication can remove the virus from an infected person's body completely (SFAF, no date). However, ART can stop the virus from replicating or slow the progression of HIV disease (WHO, 2016). Antiretroviral therapy reduces the amount of virus, and possibly the risk that an infected person will infect others (WHO, 2012). Because of effective ART, HIV

infection has changed from a disease with high morbidity and mortality to a chronic disease in developed countries such as Canada (Hogg et al., 1999).

CHAPTER 3

STATISTICAL METHODS

3.1 Introduction

In general, the objectives of joint modeling include to (i) accomplish inference for the survival outcome while considering the impact of longitudinal outcome (Brown and Ibrahim, 2003; Faucett and Thomas, 1996; Wang and Taylor, 2001; Wulfsohn and Tsiatis, 1997) and (ii) evaluate impacts on both longitudinal and survival outcomes jointly (Guo and Carlin, 2004; Henderson et al., 2000; Zeng and Cai, 2005). In a usual joint model setting, we consider a mixed-effects model for the longitudinal outcome (also regarded as longitudinal submodel) and a Cox model for the survival outcome (also regarded as survival submodel), some random effects or variables are shared by the two submodels (Wu et al., 2012). I discuss the analysis of longitudinal data, survival analysis, competing risks analysis, and joint modeling in Section 3.2, Section 3.3, Section 3.4, and Section 3.5, respectively.

3.2 Analysis of longitudinal data

3.2.1 Introduction

Longitudinal studies play a vital role in health sciences by taking repeated measurements from the same individual over time (Ware, 1985). In cross-sectional studies, a single outcome $y_i (i = 1, 2, \dots, n)$ and $p \times 1$ vector $\boldsymbol{x}_i$ of covariates are measured once for the i^{th} individual. In contrast, longitudinal data are comprised of repeated observations $y_{ij} (i = 1, 2, \dots, n; j =$

$1, 2, \ldots, m_i$) and $p \times 1$ vector $\boldsymbol{x}_{ij}$ of covariates over time t_{ij} for i^{th} subject at j^{th} time (Liang and Zeger, 1986). Here n indicates the total number of individuals in a study and m_i indicates the number of measurements for the i^{th} individual. We can collect longitudinal data either prospectively, or retrospectively from person's previous medical records (Diggle, Heagetry, Liang, and Zeger, 2002).

Since repeated measurements from the same individual can be correlated, we need special statistical techniques to analyze longitudinal data (Fitzmaurice, Laird, and Ware, 2011). Two important sources of variability in longitudinal data that may influence directly the correlation among the repeated observations are between-subject variation and within-subject variation (Fitzmaurice et al., 2011).

Missing data is common in longitudinal studies (Ibrahim and Molenberghs, 2009). For example, all study participants may not come for a scheduled visit; some participants may quit the study before ending of the study (Fitzmaurice et al., 2011). Since all participants may not have the same number of repeated observations at a common time point, the data can be unbalanced over time (Fitzmaurice et al., 2011). Therefore, for analyzing longitudinal data, we should apply techniques that can manage unbalanced and possibly incomplete data (Fitzmaurice et al., 2011).

3.2.2 Modeling longitudinal data

When modeling longitudinal data, our interest is to study the association between a response/dependent variable and a set of explanatory/independent variables (Liang and Zeger,

1986; SAS, no date). The response variables can have either continuous (e.g. Gaussian) outcomes or discrete (e.g. binary, count) outcomes (Duchateau, Janssen, and Rowlands, 1998).

Let y_{ij} indicate the response variable of the i^{th} person $(i = 1, 2, \dots, n)$ for the j^{th} measurement $(j = 1, 2, \dots, m_i)$ (Fitzmaurice et al., 2011). Since individuals may have a different number of repeated measures and may be measured at a different set of occasions, we assume that m_i repeated measurements are available for the i^{th} individual and that we observe each y_{ij} at time t_{ij} (Fitzmaurice et al., 2011). For convenience, the m_i repeated measures of the response variable for the i^{th} individual can be grouped into an $m_i \times 1$ vector such that (Fitzmaurice et al., 2011):

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{im_i} \end{bmatrix}, \qquad i = 1, 2, \dots, n.$$

Here $\boldsymbol{y}_i$ is a vector of the m_i responses over time for the i^{th} individual. The vectors of the n individuals are assumed to be independent of each other (Fitzmaurice et al., 2011). However, the repeated measures obtained from the same individual are assumed to be dependent (Fitzmaurice et al., 2011).

Let $\boldsymbol{x}_{ij}$ be a $p \times 1$ vector of covariates associated with y_{ij} (Fitzmaurice et al., 2011):

$$\boldsymbol{x}_{ij} = \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ \vdots \\ x_{ijp} \end{bmatrix}, \qquad i = 1, 2, \dots, n;\ j = 1, 2, \dots, m_i.$$

Each row of the $\boldsymbol{x}_{ij}$ vector indicates one covariate. The covariates in $\boldsymbol{x}_{ij}$ can be fixed (e.g. gender) and/or time dependent (e.g. smoking status). The vector $\boldsymbol{x}_{ij}$ can be grouped into an $m_i \times p$ matrix as follows (Fitzmaurice et al., 2011):

$$X_i = \begin{bmatrix} \boldsymbol{x}'_{i1} \\ \boldsymbol{x}'_{i2} \\ \vdots \\ \boldsymbol{x}'_{im_i} \end{bmatrix}, \qquad i = 1, 2, \dots, n,$$

where $\boldsymbol{x}'_{ij}$ indicates the *transpose* of $\boldsymbol{x}_{ij}$. The matrix X_i has the form (Fitzmaurice et al., 2011):

$$X_i = \begin{bmatrix} x_{i11} & x_{i12} & \dots & x_{i1p} \\ x_{i21} & x_{i22} & \dots & x_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{im_i1} & x_{im_i2} & \dots & x_{im_ip} \end{bmatrix},$$

where the columns indicate the p distinct covariates and the rows indicate the covariates associated with the responses at the m_i different measurement times. Let $\boldsymbol{\varepsilon}_i$ be an $m_i \times 1$ vector of random errors related with the corresponding components of the vector of responses on the i^{th} subject (Fitzmaurice et al., 2011):

$$\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{im_i} \end{bmatrix}, i = 1, 2, \dots, n.$$

The main objective of longitudinal data analysis is to model the expected value $E(\boldsymbol{y}_i)$ of the dependent/response variable $\boldsymbol{y}_i$ as a linear/nonlinear function of the independent/explanatory variables (SAS, no date). Since measurements on the same subjects are likely to be correlated, statistical analysis must incorporate correlation between measurements within the same

subject (Fitzmaurice et al., 2011; Littell, Pendergast, and Natarajan, 2000). We can do this by modeling the covariance or correlation structure of each subject's response (Fitzmaurice and Ravichandran, 2008). When data is missing, it is necessary to model covariance or correlation correctly to get accurate estimates of the regression parameters (Fitzmaurice et al., 2011). Common approaches for modeling the covariance structure include unstructured covariance model, covariance pattern models, and random effects or mixed effects models (Fitzmaurice et al., 2011). An in-depth discussion of unstructured covariance model and covariance pattern models can be found in Fitzmaurice et al. (2011). Mixed effects models are discussed in Section 3.2.2.1.

3.2.2.1 Linear mixed effects model

In a linear mixed or random effects model, we assume that the dependent variable is a linear function of independent variables with regression coefficients that vary randomly from one person to another (Diggle et al., 2002). This variation among individuals arises because of unmeasured factors (Diggle et al., 2002). Each person in the population is supposed to have his/her own mean response trajectories (Finucane, Samet, and Horton, 2007; Fitzmaurice and Ravichandran, 2008). Linear mixed effects models are commonly used for analyzing longitudinal Gaussian data (SAS, no date). The term 'mixed' is used because linear mixed effects models include both fixed and random effects (Fitzmaurice and Ravichandran, 2008). In linear mixed effects models, the mean response is modeled usually using both fixed and random effects (Bates, Mächler, Bolker, and Walker, 2015). Random effects are subject-specific, and fixed

effects are assumed to be common for all individuals (Fitzmaurice, Davidian, Verbeke, and Molenberghs, 2008).

The general linear mixed effects model (Harville, 1977; Laird and Ware, 1982; Verbeke and Molenberghs, 2000) with additional predictors can be written as (Fitzmaurice et al., 2011):

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \qquad (3.1)$$

where

$\boldsymbol{y}_i = (y_{i1}, y_{i2}, \dots, y_{im_i})'$ is the $(m_i \times 1)$ vector of repeated measurements for subject i,

$\boldsymbol{\beta}$ is a $(p \times 1)$ vector for fixed effects parameters,

X_i is a $(m_i \times p)$ known design matrix corresponding to fixed effects $\boldsymbol{\beta}$,

$\boldsymbol{b_i}$ is a $(q \times 1)$ vector of random effects parameters that completes the characterization of between-subject variation,

Z_i is a $(m_i \times q)$ known design matrix corresponding to random effects $\boldsymbol{b}_i$, with $q \leq p$, and

$\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{im_i})'$ is the $(m_i \times 1)$ vector of measurement or sampling errors that completes the characterization of within-subject variation.

We assume that the random effects $\boldsymbol{b}_i$ follow a multivariate normal distribution such that (Fitzmaurice et al., 2011):

$$\boldsymbol{b}_i \sim N(0, G),$$

where G is a covariance matrix for $\boldsymbol{b}_i$. We also assume that sampling errors, $\boldsymbol{\varepsilon}_i$, are independent of $\boldsymbol{b}_i$, i.e., $\mathrm{Cov}(\boldsymbol{b}_i, \boldsymbol{\varepsilon}_i) = 0$, and have a multivariate normal distribution (Fitzmaurice et al., 2011):

$$\boldsymbol{\varepsilon}_i \sim N(0, R_i).$$

Therefore, we can write,

$$E\begin{bmatrix}\boldsymbol{b}_i \\ \boldsymbol{\varepsilon}_i\end{bmatrix} = \begin{bmatrix}0 \\ 0\end{bmatrix}, \qquad Var\begin{bmatrix}\boldsymbol{b}_i \\ \boldsymbol{\varepsilon}_i\end{bmatrix} = \begin{bmatrix}G & 0 \\ 0 & R_i\end{bmatrix},$$

where, G and R_i represent covariance matrices for $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$, respectively. We further assume that R_i represents the diagonal matrix, $\sigma^2 I_{m_i}$, with an $m_i \times m_i$ identity matrix, I_{m_i} (Fitzmaurice et al., 2011). The covariance of $\boldsymbol{y}_i$ is, therefore (Fitzmaurice et al., 2011),

$$Cov(\boldsymbol{y}_i) \equiv V_i = Cov(Z_i \boldsymbol{b}_i) + Cov(\boldsymbol{\varepsilon}_i)$$

$$= Z_i Cov(\boldsymbol{b}_i) Z_i{}' + Cov(\boldsymbol{\varepsilon}_i)$$

$$= Z_i G Z_i{}' + R_i$$

$$= Z_i G Z_i{}' + \sigma^2 I_{m_i}.$$

Since the expected values of random effects are zero, the distribution of $\boldsymbol{y}_i$ has a mean vector $X_i\boldsymbol{\beta}$ and a covariance matrix V_i (Brown and Prescott, 2006).

In mixed effects models, we not only estimate parameters for the mean response changes of the population but also predict the change of each individual's response trajectories over time (Fitzmaurice et al., 2011). That's why, these models are used in joint modeling of

longitudinal and survival data, which will further be discussed in Section 3.5 (Rizopoulos, 2012). Mixed effects models also allow imbalance in the data; it is not necessary to have the equal number of measurements on each subject from the same set of times (Rizopoulos, 2012; Fitzmaurice et al., 2011).

Estimation:

Maximum likelihood (ML) method is used to estimate parameters in the linear mixed effects model (Rizopoulos, 2012). The maximum likelihood estimator (MLE) can be obtained by maximizing the joint probability (likelihood function) for values of the data (Fitzmaurice et al., 2011). If the observations in a model are assumed to be independent, the likelihood function is simply the product of the density functions of each observation (Asimow and Maxwell, 2015). However, there are m_i repeated measures from the same individual in the mixed model (3.1). Therefore, we cannot assume that these repeated measures within an individual are independent (Fitzmaurice et al., 2011). Hence, the joint probability density function for the vector of repeated observations should be considered (Fitzmaurice et al., 2011).

Since we assume that the vectors $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \dots, y_{im_i})'$ of repeated observations in (3.1) are independent of one another, the log-likelihood function, $l(\boldsymbol{\theta})$, can be written as a sum of the individual multivariate normal probability density functions for $\boldsymbol{y}_i$ given X_i (Fitzmaurice et al., 2011). To get the MLE of $\boldsymbol{\theta}$, we have to maximize the following log-likelihood function (Fitzmaurice et al., 2011):

$$l(\boldsymbol{\theta}) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log|V_i| - \frac{1}{2}\left\{\sum_{i=1}^{n}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})'{V_i}^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})\right\}, \qquad (3.2)$$

where $\boldsymbol{\theta}$ indicates the vector of all parameters divided into the subvectors $\boldsymbol{\theta}' = (\beta', \sigma^2, \theta_b')$, with $\theta_b = vech(G)$, $N = (\sum_{i=1}^{n} m_i)$ is the total number of observations, m_i is the total number of observations for the i^{th} subject, and $|V_i|$ indicates the determinant of square matrix V_i (Rizopoulos, 2012).

We can get the MLE of $\boldsymbol{\beta}$ by equating the score function (the derivative of the log-likelihood function (3.2) with respect to $\boldsymbol{\beta}$) to zero and solving the equation (Fitzmaurice et al., 2011). Since the first two terms do not involve $\boldsymbol{\beta}$, we can ignore them to maximize the log-likelihood function (3.2) with respect to $\boldsymbol{\beta}$ (Fitzmaurice et al., 2011). Also, the third term involves a negative sign; therefore, to maximize the equation (3.2) with respect to $\boldsymbol{\beta}$, we can simply minimize (Fitzmaurice et al., 2011)

$$\sum_{i=1}^{n} (\boldsymbol{y}_i - X_i\boldsymbol{\beta})' {V_i}^{-1} (\boldsymbol{y}_i - X_i\boldsymbol{\beta}). \qquad (3.3)$$

By minimizing the above expression (3.3), we get the Generalized Least Squares (GLS) estimator of $\boldsymbol{\beta}$ as (Fitzmaurice et al., 2011)

$$\widehat{\boldsymbol{\beta}} = \left\{ \sum_{i=1}^{n} (X_i' {V_i}^{-1} X_i) \right\}^{-1} \sum_{i=1}^{n} \left(X_i' {V_i}^{-1} \boldsymbol{y}_i \right). \qquad (3.4)$$

This formulation is based on the assumption that V_i is known (Fitzmaurice et al., 2011). However, V_i is not known in general, therefore we must estimate this matrix from the data in hand (Fitzmaurice et al., 2011). We get the MLE of V_i by maximizing $l(\theta_b, \sigma^2)$ for a given value of $\boldsymbol{\beta}$ (Rizopoulos, 2012). After getting the MLE of V_i, we substitute the estimate of V_i, by $\widehat{V}_i$ in expression (3.4) to get the MLE of $\boldsymbol{\beta}$ (Fitzmaurice et al., 2011):

$$\widehat{\boldsymbol{\beta}} = \left\{\sum_{i=1}^{n} (X_i' \hat{V}_i^{-1} X_i)\right\}^{-1} \sum_{i=1}^{n} \left(X_i' \hat{V}_i^{-1} \boldsymbol{y}_i\right). \quad (3.5)$$

In large or asymptotically large samples, $\widehat{\boldsymbol{\beta}}$ has multivariate normal distribution with mean $\boldsymbol{\beta}$ and covariance (Fitzmaurice et al., 2011)

$$Cov(\widehat{\boldsymbol{\beta}}) = \left\{\sum_{i=1}^{n} (X_i' \hat{V}_i^{-1} X_i)\right\}^{-1}.$$

Restricted maximum likelihood (REML) estimation method can also be used for estimating V_i (Harville, 1977; Rizopoulos, 2012). In REML estimation of V_i, the likelihood does not contain $\boldsymbol{\beta}$ and is defined only in terms of V_i (Fitzmaurice et al., 2011; Rizopoulos, 2012).

3.3 Survival analysis

3.3.1 Introduction

In survival analysis, we analyze the time until the occurrence of a specific event (Chernick and Friis, 2003). The time is often called survival time, failure time, or event time (Rizopoulos, 2012). Example includes:

- Time until discharge from hospital
- Time until AIDS or death among HIV patients

Usually, the exact time-to-event is observed for only a proportion of the subjects/individuals in a study (Sullivan, 2012). For all others, the time to the event is greater than the available follow-up time. This phenomenon is considered as censoring, and it is very common in survival data (Sullivan, 2012). Censoring occurs mainly for three reasons (Kleinbaum and Klein, 2012):

- Subjects do not have the event by the end of the study
- Subjects are lost to follow-up during the study
- Subjects withdraw from the study

The above censoring is known as right censoring. Other types of censoring are left censoring and interval censoring. However, in this thesis, I shall focus only on right censoring.

Right-censoring can further be divided as (Rizopoulos, 2012):

- Fixed type I censoring: The follow-up time or end of the study is pre-specified. An individual who does not experience an event is censored at the end of the study.
- Random type I censoring: The follow-up time is pre-specified. An individual may move from the study. Thus, all individuals may not have the same censoring time.
- Fixed type II censoring: The study is completed after recording a pre-specified number of an event (Rizopoulos, 2012).

Censoring can also be either informative or non-informative (Rizopoulos, 2012). If a person withdraws from the study because his/her prognosis worsens, his/her censoring can be considered informative (or non-random) (Rizopoulos, 2012). The failure rate of this person can be different from those who are still in the study. Informative censoring and missing not at random (MNAR) missing data in longitudinal studies have similar nature (Rizopoulos, 2012). Censoring is non-informative if the withdrawal of a person from the study is independent of his/her prognosis (Rizopoulos, 2012).

In standard survival analysis, it is assumed that censoring is non-informative (Allison, 2010). In this survival analysis, we are interested in the survival function and the hazard function (Bewick, Cheek, and Ball, 2004).

3.3.2 Functions in survival analysis

3.3.2.1. Survival function

Let T^* be a continuous random variable for time-to-event with probability density function (pdf) $f(t)$ and cumulative distribution function (cdf) $F(t) = P(T^* \leq t)$ (Pintilie, 2006). Then the survival function $S(t)$ is defined as the probability that the event occurs after time t (Pintilie, 2006):

$$S(t) = P(T^* > t) = 1 - F(t) = \int_t^{\infty} f(x)dx$$

The properties of the survival function are (Kleinbaum and Klein, 2012):

- Survival function is non-increasing
- When $t = 0,\ S(t) = S(0) = 1$
- When $t = \infty,\ S(t) = S(\infty) = 0.$

We get the survival curve by plotting $S(t)$ against time t (Bewick et al., 2004).

3.3.2.2 Hazard function

The hazard function describes the instantaneous event rate for an individual who survives up to time t without having an event and is defined as (Pintilie, 2006):

$$h(t) = \lim_{\delta t \to 0} \{ \frac{P(t < T^* \leq t + \delta t | T^* > t)}{\delta t} \}, t > 0$$

$$= \lim_{\delta t \to 0} \{ \frac{P(t < T^* \leq t + \delta t)}{\delta t P(T^* > t)} \}$$

$$= \lim_{\delta t \to 0} \{ \frac{F(t+\delta t) - F(t)}{\delta t P(T^* > t)} \}$$

$$= \frac{f(t)}{S(t)} \tag{3.6}$$

Here, $f(t)$ indicates the density function of the random variable T^* for time-to-event. The cumulative hazard function $H(t)$ is defined as the cumulative hazard up to time t (Rizopoulos, 2012):

$$H(t) = \int_0^t h(u)du$$

If any of the functions $S(t), h(t)$, or $H(t)$ are known, the other two functions can be derived as follows (Pintilie, 2006):

$$h(t) = -\frac{\delta}{\delta t} \log\big(S(t)\big)$$

$$H(t) = -\log\big(S(t)\big)$$

$$S(t) = \exp\big(-H(t)\big) \tag{3.7}$$

$$= \exp\left\{ -\int_0^t h(u)du \right\}$$

We must consider censoring if we want to estimate hazard function or survival function or any other characteristic of the survival time distribution using a random sample (Rizopoulos, 2012). Let T_i^* and C_i indicate the event time or survival time and censoring time for the i^{th}subject, respectively (Rizopoulos, 2012). Let δ_i be the event indicator such that (Rizopoulos, 2012)

$$\delta_i = \begin{cases} 1, & if\ i^{th} subject\ experienced\ the\ event\ (T_i^* \leq C_i) \\ 0, & if\ i^{th} subject\ \ was\ censored\ (T_i^* > C_i) \end{cases}$$

Then the observed time for the i^{th}subject is (Rizopoulos, 2012)

$$T_i = \min(T_i^*, C_i).$$

In the analysis of survival data, our objective is to estimate the characteristics of the distribution of T_i^* using information $\{T_i, \delta_i\}$ (Rizopoulos, 2012).

3.3.3 Estimation of the survival function

3.3.3.1 Non-parametric method

The Kaplan-Meier (K-M) estimate (Kaplan and Meier, 1958) is the most well-known estimate for the estimation of the survival function. It is a non-parametric estimator (Kaplan and Meier, 1958). The K-M method is very popular to compare the survival curves for two or more groups of individuals (Kleinbaum and Klein, 2012). In this method, the log-rank and Wilcoxon tests are available to compare survival differences between two or more groups of individuals (Collett, 2003). This non-parametric estimate is also called the product-limit estimate (Kleinbaum and Klein, 2012).

Suppose $t_{1,}\ t_2,\ \ldots, t_n$ be the observed survival times for n subjects (Collett, 2003). The observed survival time may be same for some subjects. Consider that r subjects experienced the event, with $r \leq n$ and the r ordered failure times are $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$. Then the Kaplan-Meier estimate of the survival function $\hat{s}_{KM}(t)$ is defined as the probability of surviving through the interval from $t_{(f)}$ to $t_{(f+1)}$, and all prior intervals (Collett, 2003):

$$\hat{s}_{KM}(t) = \prod_{g=1}^{f}\left(\frac{n_g - d_g}{n_g}\right), \tag{3.8}$$

where,

n_g indicates the number of subjects who are alive just before time $t_{(g)}$ $(g = 1, 2, \dots, r)$ and are about to fail at this time,

d_g indicates the number who fail at time $t_{(g)}$,

and $t_{(f)} \leq t < t_{(f+1)}$ for $f = 1, 2, \dots, r$.

3.3.4 Modeling survival data

In survival analysis, we are mainly interested in the risk or hazard of death/failure at any time after the starting point of the study (Collett, 2003). Therefore, we model the hazard function directly in this analysis. This modeling has two broad objectives: 1) to examine which explanatory/independent variables have an impact on the hazard function, 2) to estimate the hazard function for the individual in a study (Collett, 2003).

Both parametric and semiparametric survival models are available to examine the relationship between survival time and one or more predictor (age, gender, race, etc.) (Fox, 2008). In parametric survival models, we assume that survival times follow some known probability distributions (Kleinbaum and Klein, 2012). Commonly used parametric models include the exponential, Weibull, lognormal, and Gamma distributions. In this thesis, I shall apply a semiparametric regression model for the analysis of survival data.

3.3.4.1 Cox proportional hazards (PH) model

The Cox proportional hazards regression model, proposed by Cox (1972), is a popular semiparametric regression model for the analysis of survival data (Fox and Weisberg, 2011). Using this model, we can test if survival times between two or more groups are different after adjusting for other covariates (Singh and Mukhopadhyay, 2011). Cox PH regression model for the i^{th} individual is given by (Rizopoulos, 2012):

$$h_i(t|\boldsymbol{a}_i) = h_0(t)\exp\{\boldsymbol{\gamma}'\boldsymbol{a}_i\}, \tag{3.9}$$

where

$h_0(t)$ is the unspecified baseline hazard or baseline risk function,

$\boldsymbol{a}_i' = (a_{i1}, \dots, a_{ip})$ indicates the vector for p covariates or explanatory variables,

$\boldsymbol{\gamma}$ indicates the vector of regression coefficients corresponding to $\boldsymbol{a}_i$.

The effect of covariates on the baseline hazards for an event is multiplicative in this model (Rizopoulos, 2012).

The hazards ratio for an individual i with covariate vector a_i compared to individual i' with covariate vector $a_{i'}$ is given by (Rizopoulos, 2012):

$$\frac{h_i(t|\boldsymbol{a}_i)}{h_{i'}(t|\boldsymbol{a}_{i'})} = \exp\{\boldsymbol{\gamma}'(\boldsymbol{a}_i - \boldsymbol{a}_{i'})\}.$$

The Cox regression model is called semiparametric model because the baseline hazard function is unspecified (Fox and Weisberg, 2011). To fit this model, we don't need any assumptions about the form of the baseline hazard function (Ahmed, Vos, and Holbert, 2007).

Proportional hazards models (also known as relative risk or relative hazard models) assume that the hazards ratio does not vary over time (Bewick et al., 2004). Thus it is very important to check if this 'proportionality' assumption is met in Cox regression model (Persson, 2002). The Cox regression model is extensively used in analyzing survival data and can be fitted in most statistical software packages (Fox and Weisberg, 2011).

Partial likelihood (PL) estimates are used to estimate the regression coefficients of the Cox hazards model (Cox, 1972; Lewis, 2017). Suppose that we have a data set for n individuals with r distinct failure (event) times and $n - r$ right-censored survival times (Collett, 2003). Let us consider that there is only one event at each failure time (i.e. the data does not have any ties). The estimation method needs to sort the ordered failure times, such that (Collett, 2003)

$$t_{(1)} < t_{(2)} < \ldots < t_{(r)},$$

where $t_{(g)}$indicates the g^{th} ordered failure time.

Cox (1972) formulated the required likelihood function for the PH model in equation (3.9) as (Collett, 2003):

$$L(\boldsymbol{\gamma}) = \prod_{g=1}^{r} \frac{\exp(\boldsymbol{\gamma}' \boldsymbol{a}_{(g)})}{\sum_{l \epsilon R(t_{(g)})} \exp(\boldsymbol{\gamma}' \boldsymbol{a}_l)}, \qquad (3.10)$$

where

$\boldsymbol{a}_{(g)}$ indicates the vector of covariates who fails at the g^{th} ordered failure time, $t_{(g)}$,

$R(t_{(g)})$ (risk set) indicates set of individuals who are at risk at time $t_{(g)}$.

The product in the likelihood function (3.10) is based on the individuals who have failure times. Individuals who are at risk only contribute to the denominator of the likelihood function (Collett, 2003).

Suppose $t_1, t_2, \ldots, t_n$ are the observed survival times for n individuals and δ_i indicates the event indicator such that $\delta_i = 0$, if the i^{th} survival time t_i, $i = 1, 2, \ldots, n$, is right- censored, and $\delta_i = 1$, otherwise (Collett, 2003). The likelihood function in equation (3.10) can be written in the following form (Collett, 2003):

$$L(\boldsymbol{\gamma}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\gamma}' \boldsymbol{a}_i)}{\sum_{l \epsilon R(t_i)} \exp(\boldsymbol{\gamma}' \boldsymbol{a}_l)} \right\}^{\delta_i}, \tag{3.11}$$

where $R(t_i)$ indicates the risk set at time t_i. The log-likelihood function of this equation can be expressed as (Collett, 2003)

$$logL(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{\gamma}' \boldsymbol{a}_i - log \sum_{l \epsilon R(t_i)} \exp(\boldsymbol{\gamma}' \boldsymbol{a}_l) \right\}. \tag{3.12}$$

We can get the maximum likelihood estimates of the parameter $\boldsymbol{\gamma}$ by maximizing this log-likelihood function (3.12) using numerical methods; the maximization can be done by the Newton-Raphson procedure (Collett, 2003).

3.4 Competing risks analysis

3.4.1 Introduction

Competing risks analysis is considered to be an extension of traditional survival analysis (Beyersmann, Latouche, Buchholz, and Schumacher, 2009). Competing risks occur when more

than one type of time-to-event or survival outcome is available in the study data (Kalbfleisch and Prentice, 2002; Tsiatis, 1999). For example, if the interest lies in analyzing the time to death due to cardiovascular disease, some individuals could die of other causes. Here death from other causes can be considered competing risks of death due to cardiovascular disease (Pintilie, 2007). Let us consider another example related to cancer (Pintilie, 2007). Suppose that follow-ups are done on a cohort of patients diagnosed with breast cancer. If someone is diagnosed with cancer again in the same breast, this reappearance of the cancer is known as a local recurrence (Breastcacer.Org, 2017; Pintilie, 2007). When cancer returns at another site, it is known as a regional or distant recurrence (Breastcacer.Org, 2017; Pintilie, 2007). If someone observes a regional or distant recurrence first, she could still have a subsequent local recurrence (Pintilie, 2007). When a patient with a regional or distant recurrence receives treatment, it is very likely that the treatment will modify the probability of observing the local recurrence (Pintilie, 2007). Hence, if we are interested in analyzing the time-to-event of local recurrence, all other recurrences can be considered as competing risks (Pintilie, 2007). According to Gooley et al. (1999), a competing risk is an event that either hampers the observation of the event of interest or modifies its probability of occurrence.

3.4.2 Theoretical approaches for competing risks analysis

In standard survival analysis, we use the Kaplan-Meier method and the log-rank test to estimate the cumulative incidence and to compare cumulative incidence curves, respectively (Kim, 2007; Satagopan et al., 2004). The standard Cox regression model is used to examine the impact of covariates on the event (Singh and Mukhopadhyay, 2011). However, in competing

risks scenarios, these standard methods could provide incorrect results (Kim, 2007). In general, one minus Kaplan-Meier estimate of survival function $(1 - \hat{s}_{KM}(t))$ overestimates the cumulative incidence for cause e (Bakoyanins and Touloumi, 2010). We can use this estimate in the hypothetical situation where failures from other causes have been removed, and the cause-specific hazard of interest is not affected by this removal (i.e., failure times for the different causes are independent) (Bakoyanins and Touloumi, 2010; Gaynor et al., 1993). The analysis of competing risks data involves the joint distribution of two observable random variables: failure time T^* and cause of failure D (Bakoyannis and Touloumi, no date; Kalbfleisch and Prentice, 2002; Prentice et al., 1978; Putter et al., 2007; Tisiatis, 1999).

Joint distribution of failure time and cause of failure:

The cumulative incidence function (CIF), or subdistribution, for a failure of cause e $(e = 1, \dots, K)$ can be defined as the joint probability of failure from cause e until time t in the presence of all other possible causes (Bakoyannis and Touloumi, no date; Pintilie, 2006):

$$I_e(t) = F_e(t) = P(T^* \leq t, D = e).$$

The probability that a failure of any cause occurs until time t is called the overall distribution function and is equal to the sum of the CIFs, for all causes (Pintilie, 2006):

$$F(t) = P(T^* \leq t) = \sum_{e=1}^{K} P(T^* \leq t, D = e) = \sum_{e=1}^{K} F_e\,(t).$$

Since $\lim_{t \to \infty} F_e\,(t) = P(D = e)$, the CIF for cause e is considered as "subdistribution" (Bakoyannis and Touloumi, no date). The CIFs are determined by the cause-specific hazard function

(Hinchliffe and Lambert, 2013). We define the e^{th} cause-specific hazard function at time t as (Pintilie, 2006):

$$h_e(t) = \lim_{\delta t \to 0}\{\frac{P(t \le T^* < t + \delta t, D = e | T^* \ge t)}{\delta t}\}, \qquad e = 1, \dots, K \tag{3.13}$$

$$= \lim_{\delta t \to 0}\{\frac{P(t \le T^* < t + \delta t, D = e)}{\delta t P(T^* \ge t)}\}$$

$$= \{P(T^* \ge t)\}^{-1} \lim_{\delta t \to 0}\{\frac{P(t \le T^* < t + \delta t, D = e)}{\delta t}\}$$

$$= \frac{f_e(t)}{S(t)}.$$

Here, $S(t)$ is the overall survival probability (Satagopan et al., 2004). We can get the overall hazard function $h(t)$ of any cause of failure by summing over all cause-specific hazard functions (Pintilie, 2006):

$$h(t) = \sum_{e=1}^{K} h_e(t)$$

The cumulative incidence functions $I_e(t)$ depend on the cause-specific hazard of the respective event/failure and on the total hazard through the relation (Bakoyannis and Touloumi, 2010):

$$I_e(t) = P(T^* \le t, D = e) = \int_0^t h_e(u) S(u) du$$

$$= \int_0^t h_e(u) \exp\left[-\int_0^u \sum_{e=1}^{K} h_e(v) dv\right] du. \tag{3.14}$$

3.4.3 Modeling competing risk

In competing risks data, failure can occur due to distinct and exclusive causes (Latouche et al., 2007). In analyzing this data, two methods are mostly used to study the impact of a covariate on a particular cause of failure (Andersen, Abildstrom, and Rosthoj, 2002). In the most common method, the cause-specific hazard of a failure is modeled using a Cox proportional hazards model (Cox, 1972; Prentice et al., 1978). In the second method, we model the hazard function associated with the CIF (Fine and Grey, 1999; Gray, 1988; Klein and Andersen, 2005; Pepe, 1991). The model, based on the hazard function related with the CIF, is called a proportional subdistribution hazards model (Fine, 2001; Fine and Grey, 1999). Both Cox and subdistribution hazards models assume proportional hazards but for two different quantities (Latouche et al., 2007). For the main and competing events, if the proportionality assumption is met in a Cox model, the assumption may not be valid in a subdistribution hazards model (Fine-Gray model) and vice versa (Beyersmann and Schumacher, 2007; Latouche et al., 2013). However, Grambauer, Schumacher, and Beyersmann (2010) showed that if the subdistribution hazards model is misspecified, it may still provide a summary analysis in terms of a time-averaged subdistribution hazards ratio.

Since cumulative incidence is a function of the cause-specific hazards for all the possible causes of failure as expressed in (3.14), the effects of a covariate on a cause of failure may be different in the cause-specific hazards and the subdistribution hazards models (Bakoyannis and Touloumi, no date; Latouche et al., 2013).

The cause-specific hazard models may be suitable if our research objective is related to the cause of diseases; however, subdistribution hazards models is beneficial for clinical prediction,

and for distributing resources (Austin et al., 2016; Lau et al., 2009). Subdistribution hazard model is also appropriate for evaluating actual risks of an event, and for decision making (Koller, Raatz, Steyerberg, and Wolbers, 2011; Austin et al., 2016).

3.4.3.1 Cause-Specific Hazards (CSH) approach

Cause-specific hazard for a particular cause of failure is defined as the instantaneous failure rate of this cause while other causes of failure could be present (Bakoyannis and Touloumi, no date). To study the impact of covariates on a specific cause of failure in the presence of all other possible cases, we can choose a competing risks model equivalent to the Cox proportional hazards model (Bakoyannis and Touloumi, no date; Holt, 1978). We can write a semiparametric proportional hazards model for the cause-specific hazard function as (Bakoyannis and Touloumi, no date):

$$h_e(t; \boldsymbol{a}) = h_{e0}(t) \exp(\boldsymbol{\gamma}_e' \boldsymbol{a}), \quad e = 1, 2, \dots, K \qquad (3.15)$$

where e indicates the cause of failure, $h_{e0}(t)$ represents the baseline hazard function, $\boldsymbol{\gamma}_e$ is the vector of regression coefficients, and $\boldsymbol{a}$ is a vector of covariates for the e^{th} cause. To fit this model, we can apply standard Cox regression by censoring observations which fail due to other causes except e (Cox, 1972; Szychowski, Roth, Clay, and Mittelman, 2010). Parameters in the model (3.15) can be estimated using the partial likelihood function in equation (3.11), discussed in Section 3.3.4.1.

3.4.3.2 Cumulative incidence approach

As an instantaneous rate of occurrence of a given event, cause-specific hazard functions do not quantify the overall impact of a covariate to the patient's survival (Bakoyannis and Touloumi, 2010). For a particular cause of failure, a covariate can have different effects on the cause-specific hazard and the cumulative incidence (Gray, 1988; Latouche et al., 2013). Thus, modeling the cumulative incidence may be more appropriate in some circumstances (Koller at al., 2011; Lau et al., 2009). Fine and Gray (1999) developed a method for regression modeling with CIFs (Kuk and Varadhan, 2013). Their technique uses the subdistribution hazard (Bakoyannis and Touloumi, no date; Gray, 1988). The subdistribution hazard can be expressed as a function of the cumulative incidence (Bakoyannis and Touloumi, 2010; Scrucca, Santucci, and Aversa, 2010):

$$h_e^*(t) = \lim_{\delta t \to 0}\{\frac{P[t \le T^* < t + \delta t, D = e | T^* \ge t \ \cup \ (T^* < t \cap D \ne e)]}{\delta t}\}$$

$$= \frac{f_e(t)}{1 - F_e(t)} = -\frac{dlog\{1 - F_e(t)\}}{dt}, \qquad e = 1, \dots, K \tag{3.16}$$

The equation (3.16) provides the CIF as:

$$F_e(t) = 1 - \exp\left\{-\int_0^t h_e^*(u)\, du\right\}.$$

For the subdistribution hazard, Fine and Gray (1999) suggested a Cox-type semiparametric proportional hazards model of the form (Bakoyannis and Touloumi, 2010; Scrucca et al., 2010):

$$h_e^*(t; \boldsymbol{a}) = h_{e0}^*(t) \exp\{\boldsymbol{\gamma}_e' \boldsymbol{a}\}. \tag{3.17}$$

The CIF of the above model has the form (Bakoyannis and Touloumi, 2010):

$$F_e(t;\boldsymbol{a}) = 1 - \exp\left\{-exp(\boldsymbol{\gamma}_e'\boldsymbol{a}) \int_0^t h_{e0}^*(u)\, du\right\}. \tag{3.18}$$

The difference between the cause-specific hazard (3.13) and subdistribution hazard (3.16) arises because of their risk sets (Bakoyannis and Touloumi, 2010). In the cause-specific hazard, individuals who have failed from a cause other than e before time t, are censored at the time of their failures (Kim, 2007; Lau et al., 2009). In subdistribution hazard, individuals who have failed from a cause other than e before time t, are not censored but included in the risk set for all future failures (Bakoyannis and Touloumi, 2010; Haller, Schmidt, and Ulm, 2012; Lau et al., 2009).

Estimation:

Parameters in the subdistribution hazard model are estimated based on the right censoring mechanism (Bakoyannis and Touloumi, 2010). Suppose in a study, censoring occurs because of the administrative termination of the study. Hence, we know the censoring time even for participants who experience an event before the administrative censoring time (Bakoyannis and Touloumi, 2010). We can treat individuals with failures from causes other than e as censored and replace their failure time with their administrative censoring time (Bakoyannis and Touloumi, no date). We can then fit the standard Cox proportional hazards model in the resulting dataset to estimate the parameters of $F_e(t;\boldsymbol{a})$ in equation (3.18) (Bakoyannis and Touloumi, no date).

Consider data from studies where censoring occurs from the usual random right censoring (Bakoyannis and Touloumi, no date). In this situation, we do not know the potential censoring time for a person who has experienced a competing event but remains in the risk set for cause e (the event of interest) (Bakoyannis and Touloumi, 2010). The inverse probability of censoring weighting (IPCW) technique was suggested by Fine and Gray (1999) to fit the subdistribution hazards model in this scenario.

3.5 Joint modeling

3.5.1 Introduction

A typical approach in joint modeling literature is to consider a survival submodel with measurement errors in time-varying covariates (Diaz, 2014; Ibrahim et al. 2010; Rizopoulos, 2010; Wu et al., 2012; Wulfsohn and Tsiatis, 1997). In this approach, we usually use a linear mixed effect (LME) submodel to model time-varying covariate to address measurement errors; and a Cox proportional hazards (PH) submodel to model the survival data (Wu et al., 2012). We are mainly interested in the survival process in this approach (Diaz, 2014; Rizopoulos, 2010). Rizopoulos discussed in detail about this setting in his book (Rizopoulos, 2012, Chapter 4). I focus on this approach in this thesis.

3.5.2 Survival submodel

Let T_i^* be the true event time (survival time) for individual $i, i = 1, \dots, n$ and T_i be the observed survival time (Rizopoulos, 2012). If a person does not have any event during the study time, his/her time-to-event can be *right censored* (Wu et al., 2012). For subject i, let C_i

indicates the potential censoring time, $\delta_i = I(T_i^* \leq C_i)$ indicates the event indicator with $\delta_i = 0$ when event time is right censored and $\delta_i = 1$ otherwise (Rizopoulos, 2012; Wu et al., 2012). Thus, the observed survival data for the ith individual can be written by $\{(T_i, \delta_i), i = 1,2, \dots, n\}$, where T_i is the minimum of T_i^* and C_i (Wu et al., 2012). For the i^{th} individual, let $y_i(t)$ be the observed value at time t for the internal time-dependent covariate (such as, CD4+ count) (Rizopoulos, 2012). Usually, we observe $y_i(t)$ at some specific occasions t_{ij} where measurements are taken (Rizopoulos, 2012). Hence we get the observed longitudinal outcome for subject i at time t_{ij}, $i = 1, 2, \dots, n;\ j = 1, 2, \dots, m_i$ from the measurements y_{ij} (Rizopoulos, 2012).

Let $m_i(t)$ indicates the true longitudinal outcome at time t for the i^{th}person (Rizopoulos, 2012). The true value $m_i(t)$ is unobserved and different from $y_i(t)$ as the later can have measurement error (Rizopoulos, 2012). To examine the association between $m_i(t)$ and the hazard for an event, consider the following proportional hazards model (Rizopoulos, 2012):

$$h_i(t\,|\mathcal{M}_i(t),\ \boldsymbol{a}_i) = \lim_{\delta t \to 0}\{\frac{P[t \leq T_i^* < t + \delta t | T_i^* \geq t, \mathcal{M}_i(t),\ \boldsymbol{a}_i]}{\delta t}\}$$

$$= h_0(t)\exp\{\boldsymbol{\gamma}'\boldsymbol{a}_i + \alpha m_i(t)\}, \qquad t > 0. \tag{3.19}$$

Here, $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ indicates the true values of longitudinal covariate up to time t, $h_0(.)$ indicates the unspecified baseline hazard/risk function or the hazard for a reference individual with all covariate values 0, $\boldsymbol{a}_i$ denotes the vector of the baseline covariate, and $\boldsymbol{\gamma}$ indicates the vector of regression coefficients corresponding to the vector $\boldsymbol{a}_i$ (Rizopoulos,

2012). The impact of the longitudinal data on the survival outcome at time t is measured by the parameter α (Rizopoulos, 2012).

The hazard model (3.19) assumes that the hazard for an event at time t depends only on the present value of the time-varying covariate $m_i(t)$ (Rizopoulos, 2012). However, from the relationship between survival and cumulative hazard function, we have (Rizopoulos, 2012):

$$S_i(t|\mathcal{M}_i(t),\, a_i) = \Pr(T_i^* > t\,|\mathcal{M}_i(t),\, \boldsymbol{a}_i)$$
$$= \exp\left(-\int_0^t h_0(s)\exp\{\boldsymbol{\gamma}'\boldsymbol{a}_i + \alpha m_i(s)\}\,ds\right),$$

which indicates that the survival function is a function of the whole covariate history $\mathcal{M}_i(t)$.

3.5.3 Longitudinal submodel

We assume that the hazard function $h_i(t)$ in the hazard model (3.19) depends on the *true* longitudinal outcome $m_i(t)$ at time t (Wu et al., 2012). However, for each subject, we may have this longitudinal outcome occasionally at times $\{t_{ij}, j = 1,2,\dots,m_i\}$ with measurement errors (Rizopoulos, 2012). Therefore, to examine the impact of the longitudinal outcome to the hazard for an event, we need to estimate $m_i(t)$ for each individual (Rizopoulos, 2012). We can accomplish this by fitting a mixed effects model with the available longitudinal measurements $y_{ij} = \{y_i(t_{ij}), j = 1,2,\dots,m_i\}$ of i^{th} subject (Rizopoulos, 2012). I shall emphasize the normally distributed longitudinal outcomes and use a linear mixed effects (LME) model (Laird and Ware, 1982). With similar notations used in Section 3.2.2.1, we can write (Rizopoulos, 2012):

$$y_i(t) = \boldsymbol{x}_i'(t)\boldsymbol{\beta} + \boldsymbol{z}_i'(t)\boldsymbol{b}_i + \varepsilon_i(t)$$

$$= m_i(t) + \varepsilon_i(t), \tag{3.20}$$

$$m_i(t) = \boldsymbol{x}_i'(t)\boldsymbol{\beta} + \boldsymbol{z}_i'(t)\boldsymbol{b}_i,$$

$$\boldsymbol{b}_i \sim N(0, G), \qquad \varepsilon_i(t) \sim N(0, \sigma^2),$$

where $\boldsymbol{x_i}(t)$ and $\boldsymbol{z}_i(t)$ are the design vectors for the fixed effects $\boldsymbol{\beta}$, and for the random effects $\boldsymbol{b}_i$, respectively, $\varepsilon_i(t)$ are the error terms. The random effects $\boldsymbol{b}_i$ follow a multivariate normal distribution with covariance matrix G (Fitzmaurice et al., 2011). The error terms are mutually independent, normally distributed, and independent of $\boldsymbol{b}_i$ (Rizopoulos, 2012).

To handle the measurement error, the observed longitudinal outcome $y_i(t)$ is expressed as the sum of the true longitudinal outcomes $m_i(t)$ and a random error term in the mixed effects model (Rizopoulos, 2012). The intuitive notions for using joint models are shown in Figure 3.1, where our objective is to associate the true value of the longitudinal marker (bottom panel) with the hazard for a survival outcome (top panel) (Rizopoulos, 2012). The Figure 3.1 is used from “Joint Models for Longitudinal and Time-to-Event Data With Applications in R” (Dimitris Rizopoulos, 2012) with author’s permission.

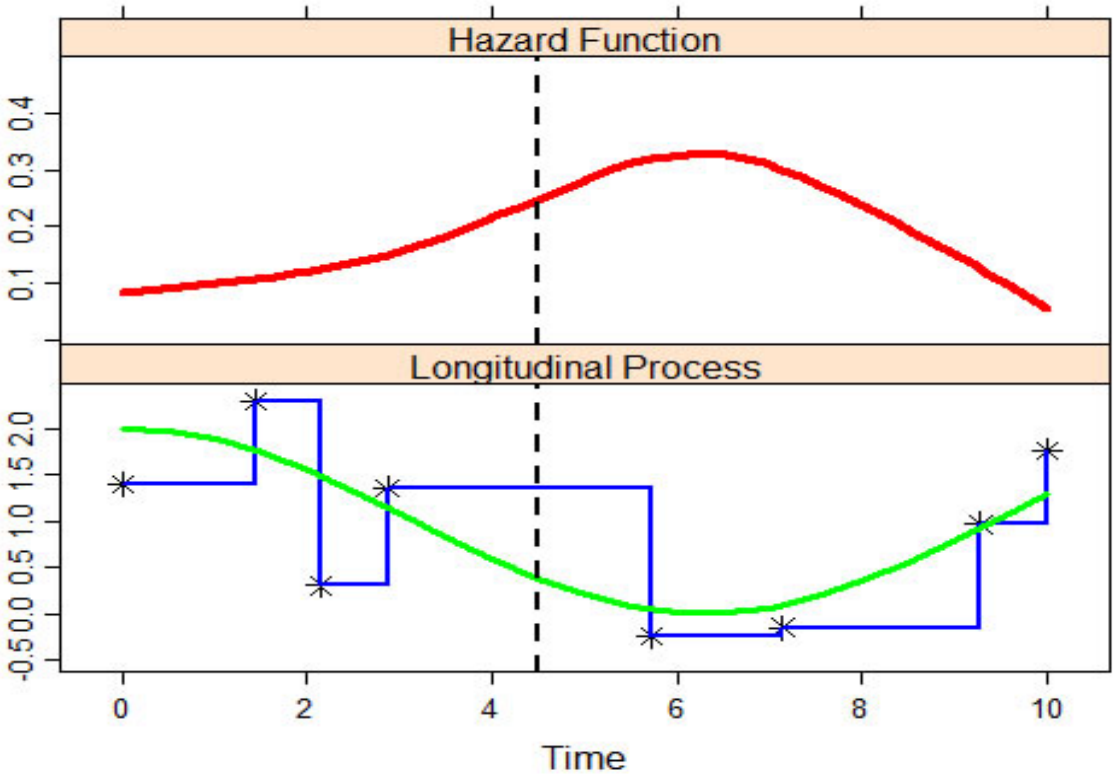


Figure 3.1: Intuitive presentation of joint models

Source: Joint Models for Longitudinal and Time-to-Event Data With Applications in R (Dimitris Rizopoulos, 2012)

The red line in the top panel shows how the hazard function changes over time (Rizopoulos, 2012). The blue line in the bottom panel corresponds to the step function provided by the extended Cox model with time-dependent covariates. The green line in the bottom panel shows the approximation of $m_i(t)$ from the joint model (Rizopoulos, 2012).

Next, I shall discuss two methods for parameter estimation in joint modeling (Wu et al., 2012):

(i) two-stage methods

(ii) likelihood methods

3.5.4 Two-stage methods

Several two-stage methods were used in joint modeling framework (Self and Pawitan, 1992; Tsiatis et al. 1995; Wu et al., 2012). Self and Pawitan (1992) used a least-square method to estimate the random effects at step one; at step two they used these estimates to assign suitable values of $m_i(t)$ which are replaced in the partial likelihood of the Cox model (Rizopoulos, 2012). To estimate parameters in the joint model, Tsiatis et al. (1995) proposed a two-stage method which combines model (3.19) with the model (3.20) (Rizopoulos, 2012).

A two-stage method works as follows (Wu et al., 2012):

Stage 1. The longitudinal covariate is modeled using a linear mixed effect (LME) model so that subject-specific values of the covariate can be estimated.

Stage 2. The survival model is fitted using the modeled values in Stage 1 as time-varying covariate values.

This two-stage modeling is better than the extended Cox model with time-dependent covariate because this approach does not use the observed longitudinal data in the survival model (Ibrahim et al., 2010). This method also reduces the bias of the parameter estimate in the Cox model (Wulfsohn and Tsiatis, 1997; Yu et al., 2004). A two-stage method is simple, and we can fit a two-stage model using existing software (Wu et al., 2012). However, this method does not use information from the longitudinal process and the survival process simultaneously in each model fitting stage (Wu et al., 2012; Yu et al., 2004). Using only the longitudinal outcome, the linear mixed effect (LME) model can provide biased estimates in the first stage

(Wu et al., 2012). As a result, estimates of the parameters in the survival model can be biased and inefficient in the second stage (Ibrahim et al., 2010).

3.5.5 Estimation by likelihood methods

The semiparametric maximum likelihood method has been used as the principal estimation method in joint modeling literatures (Henderson et al., 2000; Hsieh, Tseng, and Wang, 2006; Rizopoulos, 2012; Wu et al., 2012; Wulfsohn and Tsiatis, 1997). The joint likelihood of the longitudinal outcome and survival outcome is used in this method (Wu et al., 2012).

We get the maximum likelihood estimates from the log-likelihood function corresponding to the joint distribution of the observed data $(T_i, \delta_i, \boldsymbol{y}_i)$ (Rizopoulos, 2012). In the definition of this joint distribution, it is assumed that the association between the longitudinal and survival outcomes, and the correlation between the repeated observations are accounted by the time-independent random effects $\boldsymbol{b}_i$ (Rizopoulos, 2012). The longitudinal outcome and the survival outcome are independent given $\boldsymbol{b}_i$; hence we can write the joint distribution as (Rizopoulos, 2012)

$$f(T_i, \delta_i, \boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}) = f(T_i, \delta_i |\boldsymbol{b}_i; \boldsymbol{\theta}) f(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}), \tag{3.21}$$

and

$$f(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}) = \prod_{j=1}^{m_i} f\{y_i\left(t_{ij}\right)|\boldsymbol{b}_i; \boldsymbol{\theta}\}, \tag{3.22}$$

where $\boldsymbol{y}_i$ indicates the $m_i \times 1$ vector of longitudinal outcomes for the i^{th} subject and $\boldsymbol{\theta} = (\boldsymbol{\theta}_t', \boldsymbol{\theta}_y', \boldsymbol{\theta}_b')'$ is the vector of full parameter, with $\boldsymbol{\theta}_t$ indicating the parameters for the survival

outcome, $\boldsymbol{\theta}_y$ denoting the parameters for the longitudinal outcomes, and $\boldsymbol{\theta}_b$ indicating the parameters of the covariance matrix for the random effects (Rizopoulos, 2010; Rizopoulos, 2012).

Based on observed data of the survival and repeated measurements, the log-likelihood function for the i^{th} individual can be written as (Rizopoulos, 2012; Wulfsohn and Tsiatis, 1997):

$$log f(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}) = log \int_{-\infty}^{\infty} f(T_i, \delta_i, \boldsymbol{y}_i, \boldsymbol{b}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i \tag{3.23}$$

$$= log \int f(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) [\prod_{j=1}^{m_i} f\{y_i(t_{ij}) | \boldsymbol{b}_i; \boldsymbol{\theta}_y\}] f(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i.$$

The conditional density of the survival part $f(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta})$ can be written as (Rizopoulos, 2012; Wulfsohn and Tsiatis, 1997):

$$f(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) = h_i(T_i \mid M_i(T_i); \boldsymbol{\theta}_t, \boldsymbol{\beta})^{\delta^i} S_i(T_i \mid M_i(T_i); \boldsymbol{\theta}_t, \boldsymbol{\beta})$$

$$= [h_0(T_i) \exp\{\boldsymbol{\gamma}' \boldsymbol{a}_i + \alpha m_i(T_i)\}]^{\delta_i} \exp\left[- \int_0^{T_i} h_0(u) exp\{\boldsymbol{\gamma}' \boldsymbol{a}_i + \alpha m_i(u)\} du \right],$$

where $\boldsymbol{a}_i$ indicates the vector of baseline covariates, $\boldsymbol{\gamma}$ indicates the corresponding vector of regression coefficient, $m_i(t)$ is the true longitudinal outcome at time t, and $\boldsymbol{\beta}$ indicates the vector of fixed effects parameters in the longitudinal submodel. The impact of the longitudinal covariate to the risk for a survival outcome at time t is measured by α (Rizopoulos, 2012).

From (3.22), the joint density of the longitudinal outcomes and the random effects can be expressed as (Rizopoulos, 2012):

$$f(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta})f(\boldsymbol{b}_i;\boldsymbol{\theta}) = \prod_{j=1}^{m_i} f\{y_i\left(t_{ij}\right)|\boldsymbol{b}_i;\boldsymbol{\theta}_y\}f(\boldsymbol{b}_i;\boldsymbol{\theta}_b)$$

$$= (2\pi\sigma^2)^{-\frac{m_i}{2}}\exp\{-||\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{b}_i||^2/2\sigma^2\}$$

$$\times (2\pi)^{-q_b/2}|G|^{-1/2}\exp(-\boldsymbol{b}_i'G^{-1}\boldsymbol{b}_i/2)$$

where $||x|| = \{\sum_i x_i^2\}^{1/2}$ is the norm of the Euclidian vector and q_b indicates the dimension of the random-effects vector.

The observed data log-likelihood for all individuals in the study can be formulated as (Rizopoulos, 2012; Wulfsohn and Tsiatis, 1997)

$$l(\theta) = \sum_{i=1}^{n} logf(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}). \qquad (3.24)$$

All parameters except the baseline hazard are estimated using parametric maximum likelihood; the baseline hazard $h_0(.)$ is estimated using nonparametric maximum likelihood (Wulfsohn and Tsiatis, 1997). The baseline hazard has mass at each failure time and the number of failure times determines the dimension of this baseline hazard (Wulfsohn and Tsiatis, 1997). The Expectation-Maximization (EM) (Dempster et al., 1977) algorithm can be used for the maximization of the function (3.24) with respect to $\boldsymbol{\theta}$ (Rizopoulos, 2012). The random effects are treated as "missing data" in this EM technique (Rizopoulos, 2012). Several researchers used Gaussian quadrature rules and Monte Carlo sampling to evaluate multidimensional integrals in fitting joint models (Henderson et al., 2000; Rizopoulos, 2012; Song, Davidian, and Tsiatis, 2002; Wulfsohn and Tsiatis, 1997).

In likelihood based method, computation can be challenging because of the complicated likelihood function (Wu et al., 2012). However, this is the recommended approach for the joint analysis of longitudinal and survival data (Sweeting and Thompson, 2011).

3.5.5.1 The Expectation-Maximization (EM) algorithm

The EM algorithm is used for finding the maximum-likelihood estimates of the parameters if the data has missing values (Bilmes, 1998; Dempster et al., 1977). This algorithm is usually applied in two major situations: (i) when there are missing values in the data, and (ii) when optimizing the likelihood function is difficult (Bilmes, 1998). In practice, it is mostly used for the latter application (Bilmes, 1998). There are two steps in the EM algorithm: (i) the Expectation (E) step, and (ii) the Maximization (M) step (Rizopoulos, 2012). In the E-step, we compute the missing data using the observed data and current parameters estimates by conditional expectation; in the M-step, we maximize the conditional expectation from the first step (Bilmes, 1998; Borman, 2006).

Let $\boldsymbol{Y}^o$ and $\boldsymbol{Y}^m$ be the observed part and missing part, respectively of the complete data vector $\boldsymbol{Y}$ (Rizopoulos, 2012). For the i^{th} subject, let $\boldsymbol{\theta}^{(ic)}$ be the maximizer we estimate at iteration c ($c = 0, 1, \dots,$), and let $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(ic)})$ indicate the expectation of the joint log-likelihood (Givens and Hoeting, 2013; Rizopoulos, 2012). Using only the observed data, we want to estimate the parameters $\boldsymbol{\theta}$ for the complete data model (Rizopoulos, 2012). Each EM algorithm works as follows (Givens and Hoeting, 2013; Rizopoulos, 2012):

(i) E-step: Given the observed data, the expected value of the log-likelihood from the complete data is computed:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(ic)} = E\{logf(\boldsymbol{y};\boldsymbol{\theta})|\,\boldsymbol{y}^o;\boldsymbol{\theta}^{(ic)}\}$$

$$= \int log\, f(\boldsymbol{y}^m, \boldsymbol{y}^o;\boldsymbol{\theta}) f\left(\boldsymbol{y}^m\middle|\,\boldsymbol{y}^o;\boldsymbol{\theta}^{(ic)}\right) d\boldsymbol{y}^m,$$

(ii) M-step: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(ic)})$ with respect to $\boldsymbol{\theta}$ and set $\boldsymbol{\theta}^{(ic+1)}$ equal to the maximizer of Q.

(iii) Return to the E-step until convergence.

The EM algorithm is numerically stable (Lange, 2010; Rizopoulos, 2012). In this method, the observed data log-likelihood increases at each iteration, i.e., $logf\left(\boldsymbol{y}^o;\boldsymbol{\theta}^{(ic+1)}\right) \geq logf\left(\boldsymbol{y}^o;\boldsymbol{\theta}^{(ic)}\right)$ (Dempster et al., 1977; Rizopoulos, 2012). However, the rate of convergence can be slow if the data has a lot of missing values (Lange, 2010; Rizopoulos, 2012).

I shall use the EM algorithm to obtain the maximum likelihood estimates in the joint model discussed in Chapter 5.

E-step:

Consider the following model for i^{th} subject (Rizopoulos, 2012):

$$h_i(t) = h_0(t)\exp[\boldsymbol{\gamma}'\boldsymbol{a}_i + \alpha\{\boldsymbol{x}_i'(t)\boldsymbol{\beta} + \boldsymbol{z}_i'(t)\boldsymbol{b}_i\}],$$
$$y_i(t) = \boldsymbol{x}_i'(t)\boldsymbol{\beta} + \boldsymbol{z}_i'(t)\boldsymbol{b}_i + \varepsilon_i(t),$$
$$\boldsymbol{b}_i \sim N(0, G), \qquad \varepsilon_i(t) \sim N(0, \sigma^2),$$

$\boldsymbol{\theta} = (\boldsymbol{\theta}_t', \boldsymbol{\theta}_y', \boldsymbol{\theta}_b')'$, with $\boldsymbol{\theta}_y = (\beta',\ \sigma^2)'$, $\boldsymbol{\theta}_b = vech(G)$, and $\boldsymbol{\theta}_t = (\gamma',\ \alpha, \theta_{h_0}')'$, where θ_{h_0}indicates the parameters in the baseline hazard function $h_0(.)$. For using the EM algorithm, the random effects are considered as "missing data" (Rizopoulos, 2012). To get the parameter values $\widehat{\boldsymbol{\theta}}$ that maximize the observed data log-likelihood

$$l(\boldsymbol{\theta}) = \sum_i logf(T_i,\ \delta_i,\ \boldsymbol{y}_i;\ \boldsymbol{\theta}),$$

we maximize the expected value of the complete data log-likelihood (Rizopoulos, 2012):

$$Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(ic)}\right) = \sum_i \int_{-\infty}^{\infty} logf(T_{i,}\ \delta_i,\ \boldsymbol{y}_i,\ \boldsymbol{b}_i;\ \boldsymbol{\theta}) f(\boldsymbol{b}_i \mid T_i,\ \delta_i,\ \boldsymbol{y}_i;\ \boldsymbol{\theta}^{(ic)}) d\boldsymbol{b}_i$$

$$= \sum_i \int \left\{ logf(T_i,\ \delta_i \mid \boldsymbol{b}_i;\ \boldsymbol{\theta}_t, \boldsymbol{\beta}) + logf(\ y_i \mid \boldsymbol{b}_i;\ \boldsymbol{\theta}_{\boldsymbol{y}}) + logf(\ \boldsymbol{b}_i;\ \boldsymbol{\theta}_b) \right\}$$

$$\times f(\ \boldsymbol{b}_i \mid T_i,\ \delta_i,\ \boldsymbol{y}_i;\ \boldsymbol{\theta}^{(ic)}) d\boldsymbol{b}_i.$$

We need to apply numerical integration procedures, such as Monte Carlo sampling or Gaussian quadrature rules, to assess $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(ic)}$ (Rizopoulos, 2012).

M-step:

The log-likelihood of the complete data has three parts such that (Rizopoulos, 2012)

$$logf(T_i,\ \delta_i,\ \boldsymbol{y}_i,\ \boldsymbol{b}_i;\ \boldsymbol{\theta}) = logf(T_i,\ \delta_i \mid \boldsymbol{b}_i;\ \boldsymbol{\theta}_t, \boldsymbol{\beta}) + logf(\ \boldsymbol{y}_i \mid \boldsymbol{b}_i;\ \boldsymbol{\theta}_y) + logf(\ \boldsymbol{b}_i;\ \boldsymbol{\theta}_b).$$

Thus, for maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(ic)})$ with respect to $\boldsymbol{\theta}$, we can maximize individual parts with respect to the corresponding parameters (Rizopoulos, 2012). The variance of the measurement error and the covariance of the random effects have the expressions (Rizopoulos, 2012):

$$\hat{\sigma}^2 = \frac{\sum_i \int (\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_i)'\ (\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_i) f(\ \boldsymbol{b}_i \mid T_i,\ \delta_i,\ \boldsymbol{y}_i;\ \boldsymbol{\theta}) d\boldsymbol{b}_i}{N}$$

$$= \frac{\sum_i (\boldsymbol{y}_i - X_i\boldsymbol{\beta})'(\ \boldsymbol{y}_i - X_i\boldsymbol{\beta} - 2Z_i\widetilde{\boldsymbol{b}}_i) + trace(Z_i'Z_i\widetilde{v\boldsymbol{b}}_i) + \widetilde{\boldsymbol{b}}_i'Z_i'Z_i\widetilde{\boldsymbol{b}}_i}{N},$$

$$\hat{G} = \frac{\sum_i \widetilde{v\boldsymbol{b}}_i + \widetilde{\boldsymbol{b}}_i\widetilde{\boldsymbol{b}}_i'}{n},$$

where

$N = \sum_{i=1}^{n} m_i$, the total number of observations in the study,

$\widetilde{\boldsymbol{b}}_i = E(\,\boldsymbol{b}_i \mid T_i,\ \delta_i,\ \boldsymbol{y}_i;\ \boldsymbol{\theta}^{(it)}) = \int \boldsymbol{b}_i\, f(\,\boldsymbol{b}_i \mid T_i,\ \delta_i,\ \boldsymbol{y}_i;\ \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i$, $\widetilde{v\boldsymbol{b}}_i =$

$Variance(\,\boldsymbol{b}_i \mid T_i,\ \delta_i,\ \boldsymbol{y}_i;\ \boldsymbol{\theta}^{(it)}) = \int (\,\boldsymbol{b}_i - \widetilde{\boldsymbol{b}}_i)^2\, f(\,\boldsymbol{b}_i \mid T_i,\ \delta_i,\ \boldsymbol{y}_i;\ \boldsymbol{\theta}^{(it)}) d\boldsymbol{b}_i$.

The solutions of $\boldsymbol{\beta}$ and the parameters in the survival submodel $\boldsymbol{\theta}_t$ are obtained by Newton-Raphson update such that (Rizopoulos, 2012)

$$\widehat{\boldsymbol{\beta}}^{(it+1)} = \widehat{\boldsymbol{\beta}}^{(it)} - \left\{\frac{\delta S\left(\widehat{\boldsymbol{\beta}}^{(it)}\right)}{\delta \boldsymbol{\beta}}\right\}^{-1} S\left(\widehat{\boldsymbol{\beta}}^{(it)}\right),$$

$$\widehat{\boldsymbol{\theta}}_t^{(it+1)} = \widehat{\boldsymbol{\theta}}_t^{(it)} - \left\{\frac{\delta S\left(\widehat{\boldsymbol{\theta}}_t^{(it)}\right)}{\delta \boldsymbol{\theta}_t}\right\}^{-1} S\left(\widehat{\boldsymbol{\theta}}_t^{(it)}\right),$$

where $\widehat{\boldsymbol{\beta}}^{(it)}$ and $\widehat{\boldsymbol{\theta}}_t^{(it)}$ are the values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_t$ at the current iteration, respectively, and $\delta S\left(\widehat{\boldsymbol{\beta}}^{(it)}\right)/\delta\boldsymbol{\beta}$ and $\delta S\left(\widehat{\boldsymbol{\theta}}_t^{(it)}\right)/\,\delta\boldsymbol{\theta}_t$ are the corresponding blocks of the Hessian matrix. The elements of the score vector of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_t$ are (Rizopoulos, 2012):

$$S(\boldsymbol{\beta}) = \frac{\sum_i X_i' \left\{\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\widetilde{\boldsymbol{b}}_i\right\}}{\sigma^2} + \alpha\delta_i \boldsymbol{x}_i(T_i)$$

$$- \exp(\boldsymbol{\gamma}'\boldsymbol{a}_i) \int \int_0^{T_i} h_0\,(s)\alpha \boldsymbol{x}_i(s) \exp[\alpha\{\boldsymbol{x}_i'(s)\boldsymbol{\beta} + \boldsymbol{z}_i'(s) b_i\}]$$

$$\times f(\,\boldsymbol{b}_i \mid T_i,\ \delta_i,\ y_i; \theta) ds\ d\boldsymbol{b}_i,$$

$$S(\boldsymbol{\gamma}) = \sum_i \boldsymbol{a}_i \left[\delta_i - \exp(\boldsymbol{\gamma}' \boldsymbol{a}_i) \int \int_0^{T_i} h_0(s) \exp[\alpha\{\boldsymbol{x}_i'(s)\boldsymbol{\beta} + \boldsymbol{z}_i'(s)\boldsymbol{b}_i\}] f(\, \boldsymbol{b}_i \mid T_i,\ \delta_i,\ \boldsymbol{y}_i; \boldsymbol{\theta}) \right.$$

$$\left. \times ds\ d\boldsymbol{b}_i \right],$$

$$S(\alpha) = \sum_i \delta_i \{\boldsymbol{x}_i'(T_i)\boldsymbol{\beta} + z_i'(T_i)\widetilde{\boldsymbol{b}}_i\} - \exp(\boldsymbol{\gamma}' \boldsymbol{a}_i) \int \int_0^{T_i} h_0(s) \exp[\alpha\{\boldsymbol{x}_i'(s)\boldsymbol{\beta} + \boldsymbol{z}_i'(s)\boldsymbol{b}_i\}]$$

$$\times f(\, \boldsymbol{b}_i \mid T_i,\ \delta_i,\ \boldsymbol{y}_i; \theta) ds\ d\boldsymbol{b}_i,$$

$$S(\theta_{h_0}) = \sum_i \delta_i \frac{\delta h_0(T_i; \theta_{h_0})}{\delta \theta_{h_0}'} - \exp(\boldsymbol{\gamma}' \boldsymbol{a}_i) \int \int_0^{T_i} \frac{\delta h_0(s; \theta_{h_0})}{\delta \theta_{h_0}'} \exp[\alpha\{\boldsymbol{x}_i'(s)\boldsymbol{\beta} + \boldsymbol{z}_i'(s)\boldsymbol{b}_i\}]$$

$$\times f(\, \boldsymbol{b}_i \mid T_i,\ \delta_i,\ \boldsymbol{y}_i; \boldsymbol{\theta}) ds\ d\boldsymbol{b}_i.$$

3.5.6 Joint model diagnostics

After fitting a model, one fundamental step is to confirm the model's assumptions (Rizopoulos, 2012). To evaluate the model's assumptions, the standard tools are residual (the difference between the observed value of the response variable y and the fitted value $\hat{y}$) plots (Rizopoulos, 2012). Several articles discussed the properties and features of residuals for the separate longitudinal model and survival model (Rizopoulos, Verbeke, and Molenberghs, 2010; Therneau and Grambsch, 2000; Verbeke and Molenberghs, 2000). However, there are very limited literatures in the area of joint models diagnostics (Rizopoulos et al., 2010). Dobson and Henderson (2003) proposed conditional residuals and Rizopoulos et al. (2010) proposed the multiple imputation residuals as model assessment tools for joint models.

3.5.6.1 Residuals for longitudinal part

The subject-specific residuals and the marginal residuals are often used to assess the assumptions of the standard linear mixed-effects model (Rizopoulos et al., 2012; Verbeke and Molenberghs, 2000). We can check the validity of the assumptions of the longitudinal part of a joint model using these residuals (Rizopoulos et al., 2010). However, note that in the joint modeling framework, the residual plots only from the observed data can be deceptive (Rizopoulos et al., 2010).

If individuals drop out of the study because of the event, longitudinal measurements are missing after that point (Rizopoulos et al., 2010). In the joint modeling setting, we assume that the dropout is nonrandom (Little, 1995; Rizopoulos et al., 2010). Because of the nonrandom dropout, the observed data is not a random sample of the population (Fitzmaurice et al., 2011; Rizopoulos et al., 2010, 2012; Verbeke, Molenberghs, and Beunckens, 2008). Rizopoulos et al. (2010) proposed their method for calculating residuals and creating diagnostic plots based on random forms of the completed data set. They used multiple imputation method to compute the missing longitudinal measurements (Rubin, 1987).

3.5.6.2 Residuals for survival part

The martingale residuals are considered to be a usual type of residuals for the proportional hazards survival submodel of the joint model of the longitudinal and survival data (Barlow and Prentice, 1988; Rizopoulos, 2012; Therneau, Grambsch, and Fleming, 1990). To get the martingale residual for the i^{th}individual by time t, we need to subtract the expected number of failures obtained in the fitted model from the observed number of failures for the i^{th}individual

(Rizopoulos, 2012; Therneau et al., 1990). These residuals are mainly used to detect individuals that are not good fit by the model (Rizopoulos, 2012). The Cox-Snell residuals (Cox and Snell, 1968) are also used as residuals for the survival submodel (Rizopoulos, 2012). These residuals are the estimated values of the cumulative hazard, the negative log of the estimated survival function (Collett, 2003). If the proposed survival submodel fits data well, the Cox-Snell residuals will follow a unit exponential distribution (Collett, 2003; Rizopoulos, 2012).

CHAPTER 4

APPLICATION TO HIV STUDY

4.1 Introduction

HIV studies are a very common example in biomedical research where the longitudinal and survival outcomes have an association (Brombin, Serio, and Rancoita, 2014). Two important biomarkers longitudinally measured in HIV studies are CD4+ counts and viral load (AIDSinfo, 2014). These biomarkers are collected repeatedly over time, and survival outcomes such as AIDS, death are also recorded for each individual in HIV studies (Guo and Carlin, 2004; Lim et al., 2013; Wulfsohn and Tsiatis, 1997).

The CD4+ count is used as the strongest biomarker of HIV disease progression and the survival of HIV-infected patients (Kagan, Sanchez, Landay, and Denny, 2015). The CD4+ count decreases with the progression of HIV infection (Langford, Ananworanich, and Cooper, 2007). Usually, when an HIV-infected person starts Antiretroviral Therapy (ART), CD4+ counts increase and the amount of HIV decreases (AIDS.gov, 2016). After initiating ART, the CD4+ count is used to evaluate the impact of ART (AIDSinfo, 2014).

In this Chapter, I applied separate and joint modeling techniques discussed in Chapters 3 to a real HIV data set. The data set is described in Section 4.2. I fitted the longitudinal models in Section 4.4.1 and survival models with competing risks in Section 4.4.2. Joint models for longitudinal data and survival data with competing risks (Deslandes and Chevret, 2010) were fitted in Section 4.5.

The **JM** package in R (Rizopoulos, 2010) was used to fit joint models. Other analyses were done using SAS 9.4 (SAS Cary, NC, U.S.A.), and STATA/SE 12.1 (StataCorp. College Station, TX,

U.S.A). For this thesis, the level of significance was set at 0.05. The study was approved by the Biomedical Research Ethics Board, University of Saskatchewan (Bio # 14-314).

4.2 Data description and study population

For an application, real HIV data was obtained from the Ontario HIV Treatment Network Cohort Study (OCS). The OCS is an observational study of HIV-infected people in Ontario, Canada (McMurchy et al., 2010; Rourke et al., 2013). The purpose of the OCS is to create a prospective database to support research (Rourke et al., 2013). The OCS created the database in 1994 (Raboud et al., 2005; Rourke et al., 2013).

The OCS enrolled 5644 individuals from HIV-clinics by December 2010 (Rourke et al., 2013). At the time of enrolment, participants' average age was 41, a large proportion (60.3%) were MSM (men who have sex with men (Rourke et al., 2013). The study involves community and links data with other databases (Rourke et al., 2013). As the individuals voluntarily participated in the OCS, there is a possibility of recruitment bias in the study (Rourke et al., 2013). A detailed description of the OCS was given by Rourke et al. (2013).

HAART was widely available from 1997 (Touloumi et al., 2004). From the OCS, individuals with HIV-positive year ≥ 1997 (HAART era) were considered in our study. Demographic and clinical covariates received from the OCS included gender, age at HIV diagnosis, ethnicity, Hepatitis C Virus (HCV) infection ever, HIV risk category, ever use of ART, CD4+ count history with dates and values, viral load, AIDS-defining illness (ADI), and death. Data was released from the OCS after final approval of a formal written proposal.

There were 2345 participants from the HAART era in the OCS. Of them, 1155 individuals had CD4+ counts available at baseline (within 3 months of first HIV+ date). In this study, the survival endpoint/outcome is AIDS-Defining Illness (ADI) which will be discussed in Section 4.2.2. I mentioned in Section 1.4.4 that ADI was defined based on Opportunistic Infections (OIs) (CDC, 2008). Usually, OIs are not common among people who have CD4+ counts > 500 (AIDS.gov, 2010). Hence, to maintain homogeneity among study participants in terms of immunological characteristics, only individuals with baseline CD4+ counts less than or equal to 500 were included in the study. At baseline, 825 individuals had CD4+ counts ≤ 500. However, three of these participants had multiple survival outcomes at the same time and were excluded from further analysis. Therefore, the final study population consisted of 822 individuals (study flow chart, Figure 4.1) with a minimum of 15 years of age at diagnosis. The maximum follow-up time was 16 years. Covariate "HIV risk category" was categorized into two major groups: MSM (MSM and MSM-IDU combined) and others. I combined MSM and MSM-IDU because only 39 participants were both MSM and IDU. It would be a very small group if I would make a separate group only with the 39 participants. Participants' demographic and clinical characteristics are presented in Table 4.1 in Section 4.3.

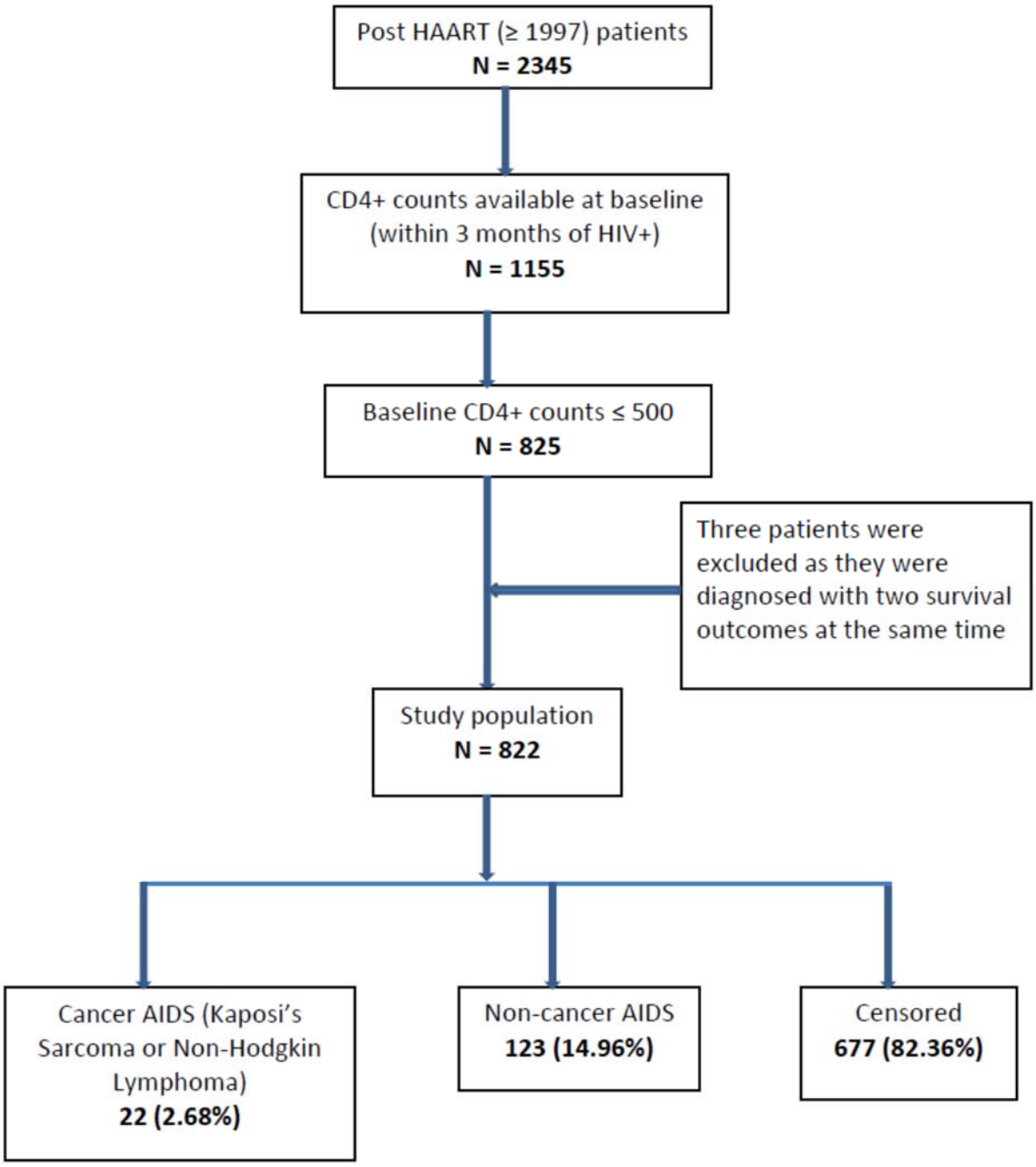


Figure 4.1: Flow chart of the study

The inclusion and exclusion criteria of our study are as follows:

Inclusion criteria:

i. HIV positive diagnosis in the Highly Active Antiretroviral Therapy (HAART) era (after 1996)

ii. HIV positive date available

iii. CD4 available at baseline (within 3 months of first HIV+ diagnosis date)

Exclusion criteria:

i. Baseline CD4+ counts greater than 500

ii. First AIDS diagnosis date at or before HIV positive date

4.2.1 CD4+ counts measurement

Individuals had large number of CD4+ count measurements because of long follow-up time of the study. The number of CD4+ count measurements also varied for individuals and the intervals of measurements were unequally spaced in the data set. Thus, I created a person-period data set for each patient covering 6-month intervals, since this biomarker is generally measured every 3 - 6 months in standard clinical practice (AIDSinfo, 2014). If a patient had two or more CD4+ counts measurements in a given interval, the average of those CD4+ counts was used for that interval. In joint model setting, maximum likelihood estimates of the parameters are obtained from the joint likelihood of the observed data. The likelihood function involves complex multiple integrals. Hence, time-space of CD4+ count measurements was fixed because of computational convenience. If CD4+ counts were not observed in any particular interval, I considered them as missing at random. For 822 participants, a total of 9115 CD4+ counts measurements were observed, with a median of 10 measurements. Individual CD4+ counts

profiles of 10 randomly selected patients are shown in Figure 4.2. Individuals were followed from the time of entry into the study until death, loss to follow-up, or censoring.

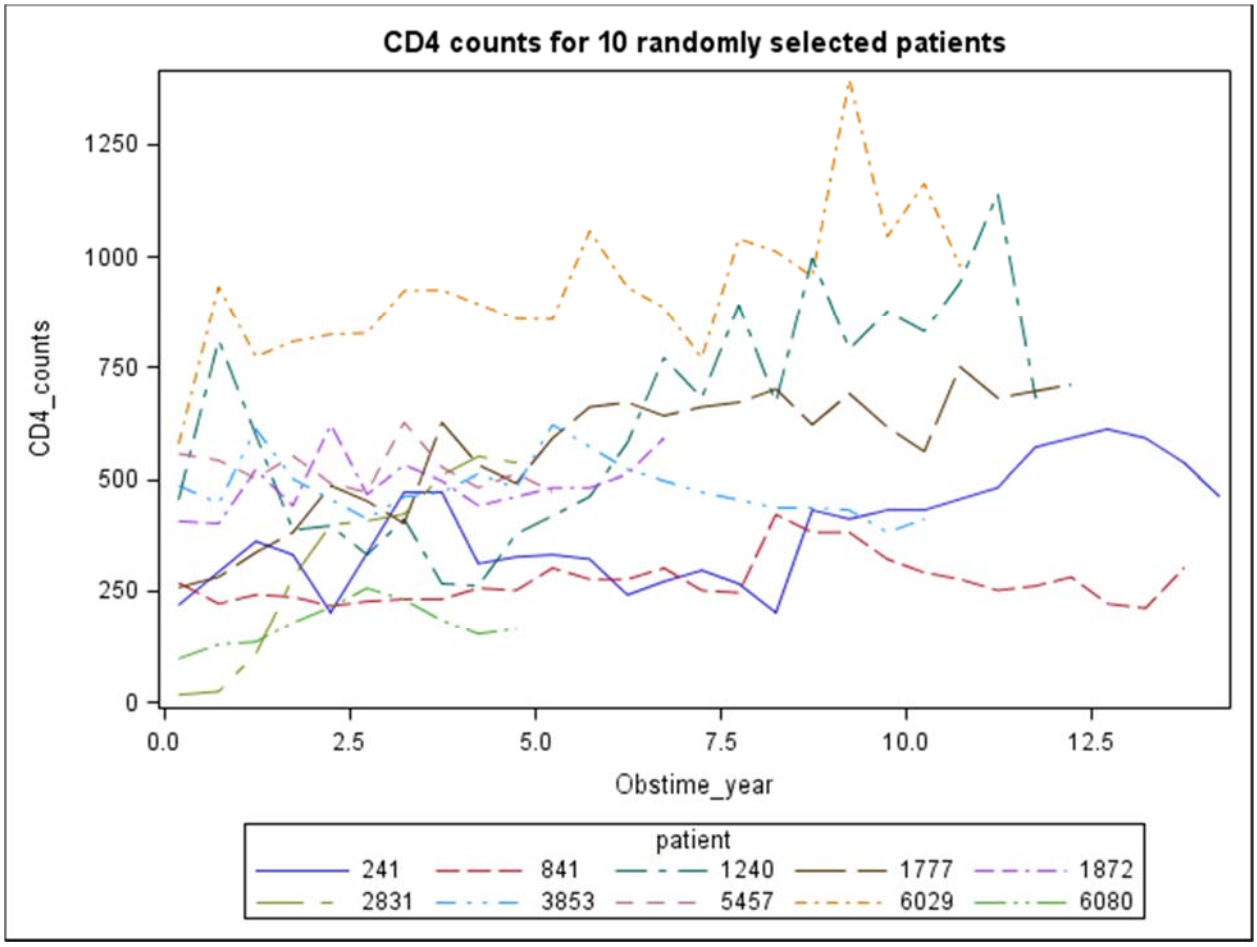


Figure 4.2: Individual CD4+ count profiles of 10 randomly selected patients

4.2.2 Survival endpoints

Cancer and HIV are associated since 1981 when Kaposi's sarcoma (KS) was reported for the first time in an immunosuppressed white MSM (Spano et al., 2008). HIV-infected people with a weak immune system were also diagnosed with Non-Hodgkin's lymphoma (NHL) and invasive cervical cancer (Ancelle-Park, 1993; Spano et al., 2008). Thus, Kaposi's sarcoma, non-Hodgkin

lymphoma, and cervical cancer are defined as AIDS-defining cancer or cancer-related AIDS (Ancelle-Park, 1993; Castro et al., 1992; Ebrahim et al., 2004). The risk of these AIDS-defining cancers is higher among HIV-infected individuals (Shiels et al., 2011). Shiels et al. (2008, 2010) considered the first AIDS-defining illness either with AIDS-defining cancer or another clinical AIDS-defining event as the event of interest/competing risk in their studies. Similarly, I also considered the first AIDS-defining illness either with AIDS-defining cancer (cancer AIDS) or another clinical AIDS-defining event (non-cancer AIDS) as the event of interest/competing event in this study. When cancer AIDS is the main event of interest then non-cancer AIDS is considered as a competing risk, and vice versa. There was no patient with cervical cancer in our study data. Hence, cancer AIDS included patients only with Kaposi's sarcoma or non-Hodgkin lymphoma. All other AIDS-defining illnesses (ADI) defined by the CDC (Section 1.4.3) were considered as non-cancer AIDS. Subsequent ADI diagnoses that happened after the first ADI diagnoses were not considered to be events in this analysis. Therefore, a patient could have either cancer AIDS or non-cancer AIDS (disjoint first event). Time-to-event was calculated from the HIV diagnosis date to the date of first ADI.

4.3 Descriptive analysis

Based on inclusion and exclusion criteria described in 4.2, a total of 822 participants diagnosed between January 1997 and October 2012 were eligible for the study. Among them, 657 (79.9%) were males, 103 (12.5%) were Hepatitis C virus-infected, and 686 (83.5%) were exposed to Anti-Retroviral (ARV) ever (Table 4.1). The majority of participants were white (55.0%), followed by Black/African (19.2%), and Aboriginal (6.8%). In the HIV risk category, the

majority was MSM (MSM and MSM-IDU, 58.5%), 17.4% had previously resided in an HIV-endemic area, and 10.3% were heterosexual. Participants' mean age was 37.4 years (SD = 10.3) at the time of HIV diagnosis. At baseline, the mean CD4+ count was 260 cells/mm^3 (SD = 145; median = 268; interquartile range = 240) and the mean log_{10} viral load was 4.5 copies/mL (SD = 0.9; median = 4.6; interquartile range = 1.1).

The median follow-up time of the study was 6.2 years (interquartile range = 7.2). Among 822 individuals, 22(2.7%) developed cancer-related AIDS and 123 (15.0%) developed non-cancer AIDS. The incidence rate of cancer AIDS was 4.2 per 1,000 person-years [95% Confidence Interval (CI): (2.7, 6.3)]. The incidence rate of non-cancer AIDS was 23.3 per 1,000 person-years [95% CI: (19.5, 27.8)].

Thirty-one (3.8%) participants of this study population died; of these, two deaths (6.5%) were HIV related, and 12 (38.7%) were non-HIV related. The cause of death for 17 participants (54.8%) was unknown (Table 4.1).

Table 4.1: Demographics and clinical characteristics of the participants (N=822)

Variable	Number (%)
Male	657 (79.9%)
Age at HIV positive date Mean (SD) Median (interquartile range)	 37.4 (10.3) 37.0 (14.0)
Race Aboriginal Multiple race Black/African South Asian White Other	 56 (6.8%) 39 (4.7%) 158 (19.2%) 40 (4.9%) 452 (55.0%) 39 (4.7%)

Unknown	38 (4.6%)
HIV risk category	
Men Sex Men (MSM)	442 (53.8%)
MSM-Injection Drug User (IDU)	39 (4.7%)
IDU	61 (7.4%)
Clotting factor	6 (0.7%)
Transfusion	7 (0.9%)
HIV-endemic	143 (17.4%)
Heterosexual transmission	85 (10.3%)
MTC mother to child transmission	1 (0.1%)
Occupational	1 (0.1%)
NIR Non-identified risk	37 (4.5%)
Ever Hepatitis C infection	103 (12.5%)
Ever ARV	686 (83.5%)
Cancer AIDS	22 (2.7%)
Kaposi's Sarcoma	18
Non-Hodgkin's Lymphoma	4
Cervical cancer	0
Non-cancer AIDS	123 (15.0%)
CD4+ counts (cells/mm^3) at baseline	
Mean (SD)	260 (145)
Median (interquartile range)	268 (240)
Log_{10} HIV viral load (copies/mL) at baseline	
Mean (SD)	4.5 (0.9)
Median (interquartile range)	4.6 (1.1)
Follow-up time, median (interquartile range)	6.2 (7.2)
Death	31
HIV-related	2 (6.5%)
Non-HIV related	12 (38.7%)
Unknown	17 (54.8%)

Men Who Have Sex With Men (MSM)

Compared with all other ethnicities, white individuals were significantly more likely to be MSM (29.9% vs. 70.1%; $P < 0.0001$) (Table 4.2). The mean baseline CD4+ count and Log_{10} HIV viral load were significantly higher for MSM compared to all other HIV risk categories (275 cells/mm^3 vs. 240 cells/mm^3; $P = 0.0008$, 4.7 copies/mL vs. 4.3 copies/mL; $P < 0.0001$ respectively). Compared to others, the mean age at the time of HIV infection among MSM was higher, although it was not significantly higher than in the other group (37.9 years vs. 36.7 years; $P = 0.12$). The proportion of ARV exposures ever were also not different between the two groups ($P = 0.38$). Nonetheless, the proportion of HCV infection ever was significantly lower among MSM (6.7% vs. 20.8%; $P < 0.0001$). Median follow-up time for MSM was significantly higher than that of the other group (6.5 vs. 5.8; $P = 0.025$).

Table 4.2: Demographic and clinical characteristics associated with MSM

Covariates	HIV risk category (MSM) (N = 481, 58.5%)	HIV risk category (Other) (N = 341, 41.5%)	*P* - value
Ethnicity (White)	70.1%	33.7%	< 0.0001†
Age in years Mean (SD) Median (IQR)	 37.9 (10.3) 37.0 (13.0)	 36.7 (10.3) 36.0 (14.0)	 0.12‡ 0.11§
Ever Hepatitis C infection	32 (6.7%)	72 (20.8%)	< 0.0001†
Ever ARV	406 (84.4%)	280 (82.1%)	0.38†
CD4+ count (cells/mm^3) at baseline Mean (SD) Median (IQR)	 275 (142) 293 (222)	 240 (146) 240 (244)	 0.0008‡ 0.0007§

Log_{10} HIV viral load (copies/mL) at baseline Mean (SD) Median (IQR)	 4.7 (0.9) 4.8 (1.1)	 4.3 (0.9) 4.4 (1.0)	 < 0.0001‡ < 0.0001§
Follow-up time, median (IQR)	6.5 (6.1)	5.8 (8.0)	0.025§

†*P*-value is based on Chi-square test; ‡*P*-value is based on T-test; §*P*-value is based on Wilcoxon test; IQR = Interquartile range

4.4 Separate analysis

4.4.1 Longitudinal analysis

To normalize CD4+ counts, I used the square root of the CD4+ counts in the models. As discussed in Section 3.2.2.1, for the analysis of the longitudinal outcome, I used linear mixed effects model with random intercept and random slope for the square root of CD4+ counts:

$$\sqrt{CD4+_{ij}} = \beta_0 + \beta_1 x_{1i} + b_{0i} + b_{1i} Time_{ij} + \varepsilon_{ij}, \qquad (4.1)$$

where $\sqrt{CD4+_{ij}}$ indicates the square root of the j^{th} CD4+ counts measurement on the i^{th} individual, $j = 1, 2, \ldots, m_i$, $i = 1, 2, \ldots, n$ and ε_{ij} is the mutually independent measurement errors.

In the univariable or unadjusted mixed effects model, follow-up time for CD4+ measurement, HIV risk category, ethnicity, and ever use of ARV were significant (Table 4.3). Age at diagnosis, gender, and ever Hepatitis C infection were not significant.

Table 4.3: Univariable mixed effects model analysis for repeated measurements

Covariates	**Estimate**	**Standard Error**	**P-value**
Time (in year)	0.79	0.03	<0.0001

HIV risk category (MSM)	1.61	0.32	<0.0001
Age at diagnosis	-0.02	0.02	0.17
Gender (Male)	0.75	0.40	0.06
Ethnicity (White)	1.12	0.32	0.0004
Ever ARV	2.55	0.44	<0.0001
Ever Hepatitis C infection	-0.17	0.48	0.71

Significant covariates in the univariable analysis were included in the adjusted or multivariable linear mixed model. Although age at diagnosis was not significant in the univariable model, it was included in the multivariable model as a potential confounder. Covariates HIV risk category, gender, and ethnicity were highly correlated. Females did not have any cancer AIDS events. Therefore, between the HIV risk category and ethnicity, the HIV risk category was included in the multivariable analyses because of the importance of exposure risk categories. Covariate "Ever ARV" indicates that a person could receive Anti-Retroviral medication anytime between his/her HIV positive date and the survival endpoint. This means that the duration of ARV treatment varied across individuals. Moreover, detailed information about ARV medications (e.g., ARV regimen) was not available in the data set. Due to these limitations, the covariate "Ever ARV" was not incorporated in the multivariable model. Finally, covariates follow-up time, HIV risk category, and age at diagnosis were included in the separate multivariable mixed effects model, as well as in the longitudinal submodel for joint modeling in Section 4.5. We fitted the following multivariable mixed effects model with random intercept and random slope for the square root of CD4+ counts:

$$\sqrt{CD4 +_{ij}} = \beta_0 + \beta_1 MSM_i + \beta_2 Age_i + \beta_3 Time_{ij} + b_{0i} + b_{1i} Time_{ij} + \varepsilon_{ij}, \qquad (4.2)$$

In the multivariable model, follow-up time and HIV risk category were significant while age at diagnosis was not significant (Table 4.4). The mean CD4+ counts were significantly higher for MSM compared to the other HIV risk category. Interaction effects between time and HIV risk category, and between age and HIV risk category were not significant.

Table 4.4: Adjusted/Multivariable mixed effects model

Covariates	**Estimate**	**Standard Error**	**P-value**
Intercept	17.23	0.59	<0.0001
Time	0.79	0.03	<0.0001
HIV risk category (MSM)	1.57	0.31	<0.0001
Age at diagnosis	-0.02	0.01	0.22

4.4.2 Survival analysis with competing risks

4.4.2.a Kaplan-Meier analysis

In the Kaplan-Meier analysis for the event of cancer AIDS, the time to cancer AIDS was significantly different between the two HIV risk categories (MSM and other) (Figure 4.3). *P*-values for Log-Rank and Wilcoxon tests were 0.032 and 0.047, respectively. Compared to MSM, the other risk group had better survival for cancer AIDS.

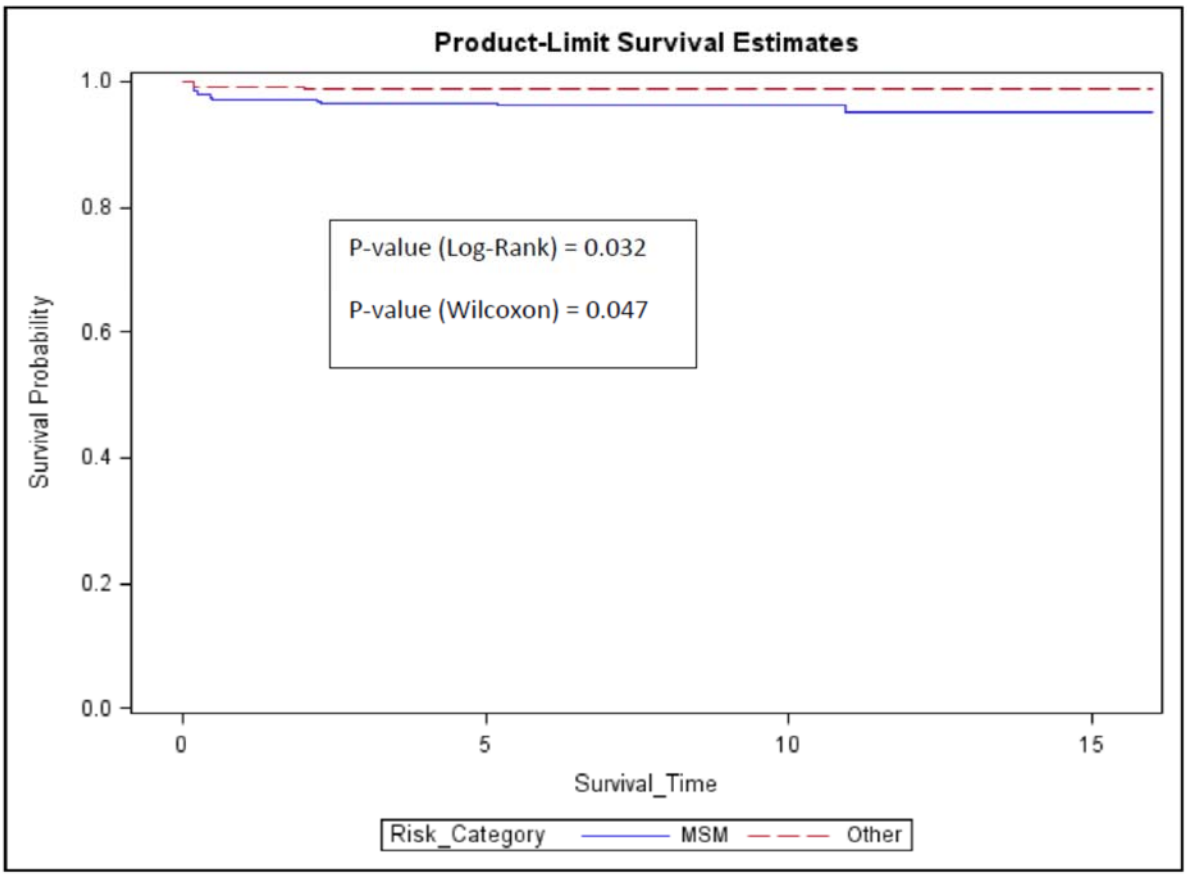


Figure 4.3: Kaplan-Meier survival plot for cancer AIDS by HIV risk category (MSM vs. Other)

The Kaplan-Meier survival curves for the length of time after HIV infection until the occurrence of non-cancer AIDS are presented in Figure 4.4. There was a significant difference in survival times between MSM and other risk groups. *P*-values for Log-Rank and Wilcoxon tests were 0.0004 and 0.0007, respectively. For non-cancer AIDS, MSM had better survival compared to the other risk group.

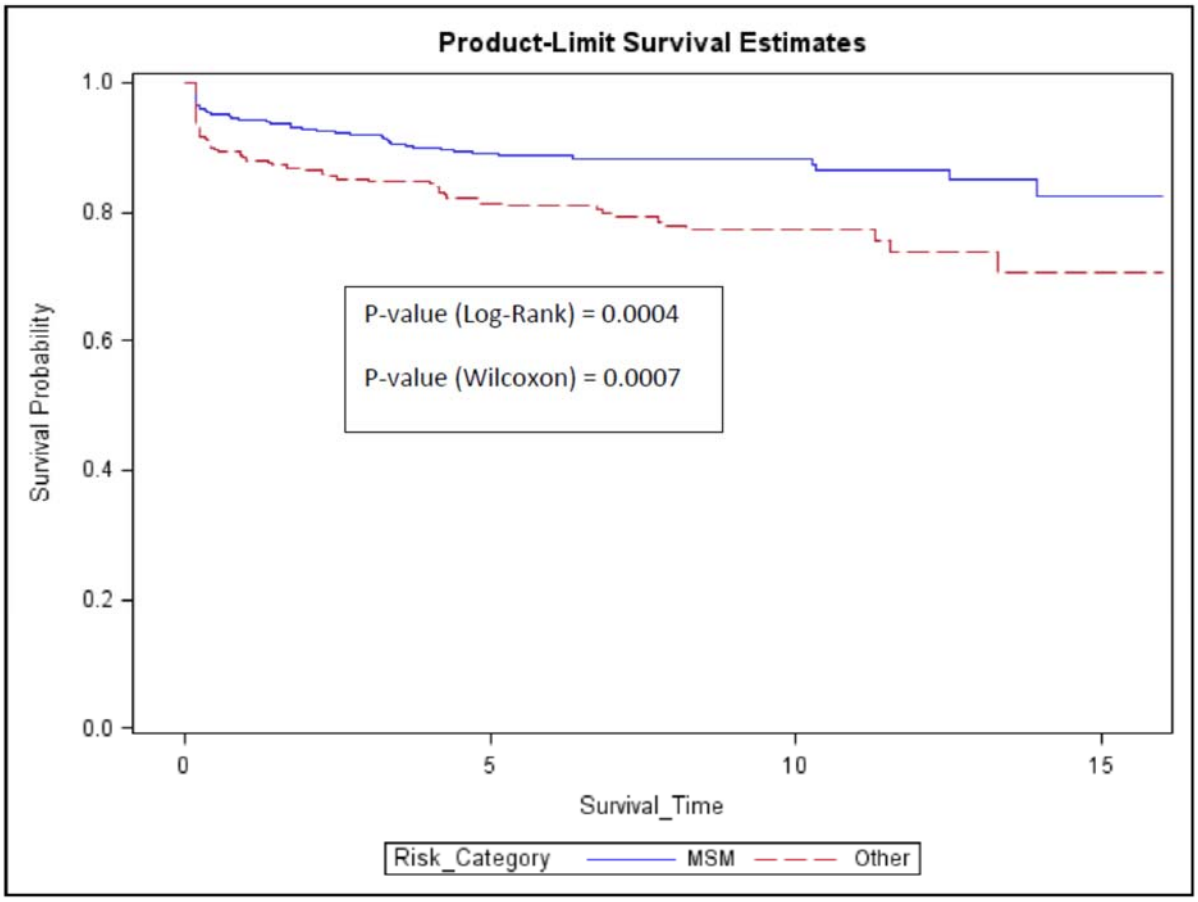


Figure 4.4: Kaplan-Meier survival plot for non-cancer AIDS by HIV risk category (MSM vs. Other)

4.4.2.b Cox Cause-Specific Hazards (CSH) model

The following univariable Cox cause-specific hazards models were applied to model cancer AIDS (risk 1) and non-cancer AIDS (risk 2), respectively:

$$h_{i1}(t) = h_{10}(t)\exp(\gamma_{11} * a_i) \qquad (4.3)$$

$$h_{i2}(t) = h_{20}(t)\exp(\gamma_{21} * a_i) \qquad (4.4)$$

where $h_{10}(t)$ and $h_{20}(t)$ are the unspecified baseline cause-specific hazards for cancer AIDS and non-cancer AIDS, respectively.

In the univariable cause-specific hazards analysis, compared to other HIV risk group, the HIV risk group MSM had significantly higher hazards for cancer AIDS but significantly lower hazards for non-cancer AIDS (Table 4.5). Older people had higher hazards for cancer AIDS. Age was not associated with non-cancer AIDS. We were not able to estimate the effect of gender for cancer AIDS as no female participants had cancer AIDS. However, males had lower hazards for non-cancer AIDS. This result is not unusual because MSM had lower hazards for non-cancer AIDS and, of course, all MSM are male. Individuals with ever use of ARV had significantly lower risk for both events. Ethnicity and hepatitis C infection were not associated with any of the events.

The number of individuals with cancer AIDS was very small in our study (n = 22; 2.7%). This might result in a relatively higher standard error (0.55) of the estimate for MSM effect on cancer AIDS and, consequently, wider 95% confidence interval of the hazards ratio.

Table 4.5: Univariable Cox cause-specific hazards model

	Cancer AIDS		**Non-cancer AIDS**	
Covariates	Estimate(SE[a])	HR[b] (95% CI[c])	Estimate(SE[a])	HR[b] (95% CI[c])
MSM	1.12 (0.55)	3.07 (1.04, 9.06)*	-0.63 (0.18)	0.53 (0.37, 0.76)*
Age at diagnosis	0.05 (0.02)	1.05 (1.02, 1.09)*	0.01 (0.01)	1.01 (1.00, 1.03)
Gender(Male)[$]			-0.62 (0.20)	0.54 (0.37, 0.79)*
White ethnicity	0.53 (0.46)	1.70 (0.69, 4.16)	-0.29 (0.18)	0.75 (0.53, 1.07)
Ever ARV	-1.92 (0.43)	0.15 (0.06, 0.34)*	-1.98 (0.18)	0.14 (0.10, 0.20)*
Ever Hepatitis C infection	-1.12 (1.02)	0.33 (0.04, 2.42)	0.09 (0.26)	1.09 (0.66, 1.83)

[a]Standard error; [b]Hazards ratio; [c]Confidence interval; *Significant at 5% level of significance; [$]Unable to estimate the effect of gender for cancer AIDS as no female participants had cancer AIDS

Although Ever ARV was significant in the univariable models, it was not included in the multivariable Cox CSH models because of the limitations of this covariate, as discussed in Section 4.4.1. Thus, only age and HIV risk category were incorporated in the multivariable or adjusted models. The following multivariable Cox CSH models were applied to model cancer AIDS (risk 1) and non-cancer AIDS (risk 2), respectively:

$$h_{i1}(t) = h_{10}(t)\exp(\gamma_{11}MSM_i + \gamma_{12}Age_i) \quad (4.5)$$

$$h_{i2}(t) = h_{20}(t)\exp(\gamma_{21}MSM_i + \gamma_{22}Age_i) \quad (4.6)$$

In the multivariable CSH model for cancer AIDS (4.5), age was significant, but MSM was not (Table 4.6). MSM was still significant in the multivariable model for non-cancer AIDS (4.6). MSM had significantly lower hazards for non-cancer AIDS [HR = 0.52; 95% CI: (0.37, 0.75)].

Table 4.6: multivariable Cox cause-specific hazards model

	Cancer AIDS		**Non-Cancer AIDS**	
Covariates	Estimate(S.E.)	HR (95% CI)	Estimate(S.E.)	HR (95% CI)
MSM	1.06 (0.55)	2.89 (0.98, 8.56)	-0.65 (0.18)	0.52 (0.37, 0.75)*
Age at diagnosis	0.05 (0.02)	1.05 (1.01, 1.09)*	0.01 (0.01)	1.01 (1.00, 1.03)

*Significant at 5% level of significance

4.4.2.c Subdistribution Hazard (SDH) model

I applied Cox-type proportional subdistribution hazards model for both events. For most of the covariates, the results in univariable subdistribution hazards models (Table 4.7) were similar but not identical to the univariable Cox CSH models (Table 4.5) for both events. However,

hazards ratio of MSM for cancer AIDS was 3.17 in SDH model, compared to 3.07 in CSH model. Hazards ratio of MSM for non-cancer AIDS was 0.53 in both the CSH and SDH models.

Table 4.7: univariable proportional subdistribution hazards model

	Cancer AIDS		**Non-Cancer AIDS**	
Covariates	Estimate(S.E.)	HR (95% CI)	Estimate(S.E.)	HR (95% CI)
MSM	1.15 (0.55)	3.17 (1.07, 9.39)*	-0.64 (0.18)	0.53 (0.37, 0.75)*
Age at diagnosis	0.05 (0.01)	1.05 (1.02, 1.08)*	0.01 (0.01)	1.01 (1.00, 1.03)
Gender (Male)[$]			-0.64 (0.19)	0.53 (0.36, 0.77)*
White ethnicity	0.55 (0.45)	1.73 (0.71, 4.21)	-0.29 (0.18)	0.75 (0.53, 1.06)
Ever ARV	-1.66 (0.42)	0.19 (0.08, 0.43)*	-1.90 (0.18)	0.15 (0.11, 0.21)*
Ever Hepatitis C infection	-1.11 (1.02)	0.33 (0.05, 2.44)	0.10 (0.25)	1.11 (0.68, 1.81)

*Significant at 5% level of significance; [$]Unable to estimate the effect of gender for cancer AIDS as no female participants had cancer AIDS

Again, only age and HIV risk category were incorporated in the multivariable SDH models. For cancer AIDS, both MSM and age were significant in the multivariable SDH model (Table 4.8). Notably, age was significant in multivariable CSH model, while MSM was not (Table 4.6). Thus, for cancer AIDS, the multivariable SDH model provided different results from the multivariable CSH model. We got similar but not identical results in both CSH and SDH multivariable models for non-cancer AIDS.

Table 4.8: Multivariable proportional subdistribution hazards model

Cancer-AIDS	**Cancer AIDS**		**Non-Cancer AIDS**	
Covariates	Estimate(S.E.)	HR (95% CI)	Estimate(S.E.)	HR (95% CI)
MSM	1.10 (0.55)	3.00 (1.02, 8.86)*	-0.66 (0.18)	0.52 (0.36, 0.73)*
Age at diagnosis	0.05 (0.01)	1.05 (1.02, 1.08)*	0.01 (0.01)	1.01 (1.00, 1.03)

*Significant at 5% level of significance

Summary of CSH and SDH analyses

For cancer AIDS, the multivariable SDH model provided different results than the multivariable CSH model. On the other hand, the multivariable CSH and SDH models provided similar results for non-cancer AIDS. When non-cancer AIDS (15.0% of all participants) was the event of interest, the proportion of the competing event (cancer AIDS) was very low (2.7% of all participants). This could be the possible reason for not getting substantially different results in the CSH and SDH models for non-cancer AIDS. MSM had significantly higher hazards for cancer AIDS but significantly lower hazards for non-cancer AIDS in multivariable SDH models (Table 4.8). Thus for MSM, hazards for cancer AIDS and non-cancer AIDS were in the opposite directions. This phenomenon highlights the importance of performing competing risks analysis, as well as considering both CSH and SDH approaches. The opposite effects of MSM on cancer AIDS and non-cancer AIDS would not have been studied if we had considered cancer AIDS and non-cancer AIDS as a combined event.

As time-to-event was calculated from HIV diagnosis date to the date of the first cancer AIDS or non-cancer AIDS, the censoring time for participants of both events was known. Hence, in

the SDH model of cancer AIDS, failure time for participants with non-cancer AIDS (competing risk) was replaced with their censoring time. Similarly, in the SDH model of non-cancer AIDS, failure time for participants of cancer AIDS was replaced with their censoring time. We then applied standard Cox proportional hazards models in the resulting datasets to estimate parameters for the subdistribution hazards models. However, we checked our results with the results from the SDH models using the inverse probability of censoring weighting technique (IPCW) (Fine and Gray, 1999). We found similar results in both techniques. The SDH model using IPCW can be fitted by PHREG procedure in current SAS software (SAS version 9.4). Previously this model could be fitted only in R and Stata.

Proportional Hazards (PH) assumption

One of the key assumptions in the Cox model relates to proportional hazards (PH). We investigated the PH assumption by testing for time-by-covariate interaction in multivariable CSH models. All models met the PH assumption. Cumulative incidence plots by HIV risk category (MSM vs. other) for cancer AIDS and non-cancer AIDS are presented in Figure 4.5 and Figure 4.6, respectively.

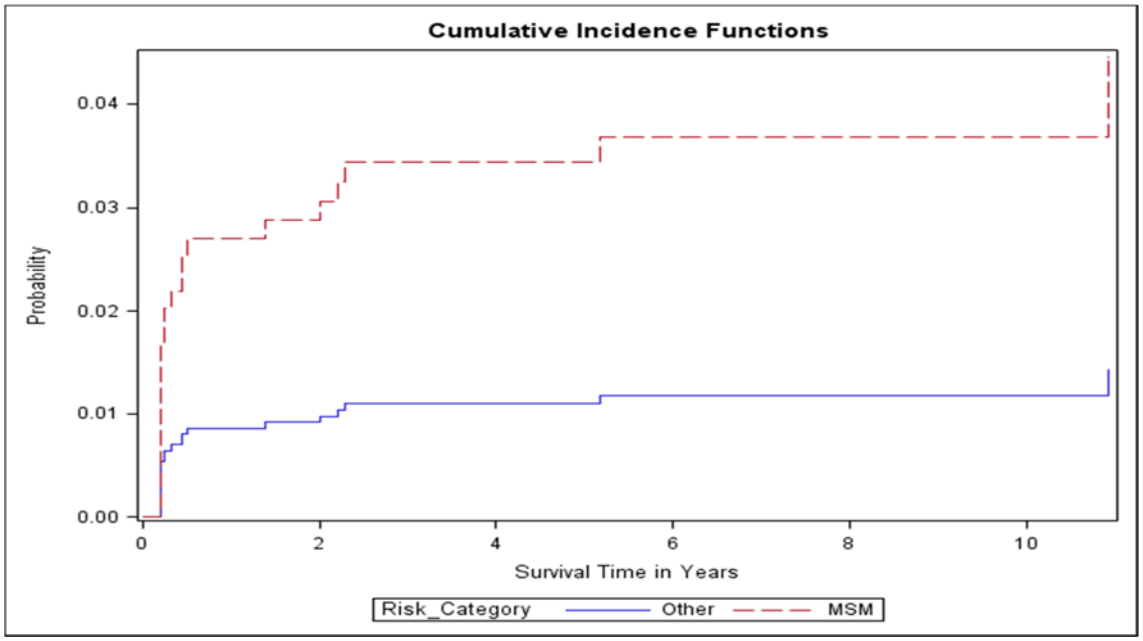


Figure 4.5: Cumulative incidence curves of cancer AIDS by HIV risk category (MSM vs. Other).

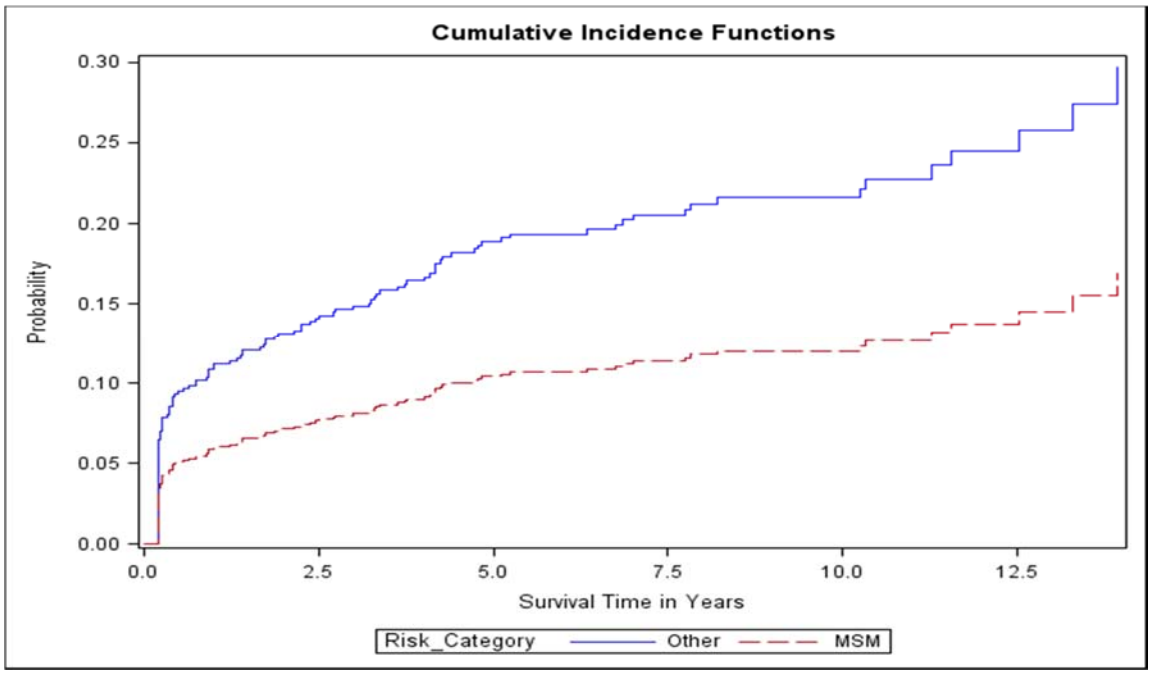


Figure 4.6: Cumulative incidence curves of non-cancer AIDS by HIV risk category (MSM vs. Other).

4.5 Joint modeling

I applied the joint modeling approach described in Chapter 3 to the OCS HIV data. Because of the limitations of the variables described in Section 4.4.1, only follow-up time, age, and HIV risk category were used in the joint modeling.

4.5.1 Longitudinal submodel

For the i^{th} subject, the following longitudinal submodel was considered for the square root of the j^{th} CD4+ counts measurement:

$$\sqrt{CD4+_{ij}} = \beta_0 + \beta_1 MSM_i + \beta_2 Age_i + \beta_3 Time_{ij} + b_{0i} + b_{1i} Time_{ij} + \varepsilon_{ij} \qquad (4.7)$$

4.5.2 Survival submodel

I used the following competing risks submodels to model cancer AIDS and non-cancer AIDS:

$$h_{i1}(t) = h_{10}(t) exp\{\gamma_{11} MSM_i + \gamma_{12} Age_i + \alpha(\beta_{10} + \beta_{11} MSM_i + \beta_{12} Age_i + \beta_{13} Time_{ij} + b_{10i} + b_{11i} Time_{ij})\}, \quad (4.8)$$

$$h_{i2}(t) = h_{20}(t) exp\{\gamma_{21} MSM_i + \gamma_{22} Age_i + \alpha(\beta_{20} + \beta_{21} MSM_i + \beta_{22} Age_i + \beta_{23} Time_{ij} + b_{20i} + b_{21i} Time_{ij})\}, \quad (4.9)$$

where $h_{10}(t)$ is the unspecified baseline cause-specific hazard of cancer AIDS and $h_{20}(t)$ is the unspecified baseline cause-specific hazard of non-cancer AIDS.

Cancer AIDS:

In the joint model of longitudinal and CSH submodels for cancer AIDS, age and MSM were significant both in longitudinal and survival submodels (Table 4.9). The mean CD4+ count was significantly higher for MSM. CD4+ counts decreased with the increment of age at diagnosis. MSM and older participants had higher hazards of cancer AIDS. The association parameter was significantly different from zero, indicating a strong association between the square root of CD4+ counts and the risk for cancer. The negative value of the association parameter (-0.21) indicated that the slope of CD4+ counts was negatively associated with the hazard for cancer AIDS, with a unit increase in this marker corresponded to a 19% decrease in the risk for cancer AIDS (HR = 0.81; 95% CI: 0.75-0.88). In the joint model of longitudinal and SDH submodels, age and MSM were significant both in longitudinal and survival submodels (Table 4.9). However, in the longitudinal submodel, the estimate of the intercept was different than the estimate of the intercept in the CSH-based joint model. CD4+ counts increased with the increment of age at diagnosis. In the survival submodel, the effect of MSM on cancer AIDS was also numerically different from the CSH-based joint model. The point estimate and corresponding 95% confidence intervals of the hazards ratio for MSM were 4.76 [1.54, 14.72] compared to 4.52 [1.51, 13.52] in the CSH-based joint model. The estimates of the association parameter were similar but not identical in the two joint models. The values of Akaike's Information Criterion (AIC) were 47376 and 54005 in the CSH-based and SDH-based joint models, respectively (Akaike, 1973). However, the values of AIC were not directly comparable because of the different survival submodels in the two joint models.

We noticed that CD4+ counts increased with the increment of age at diagnosis in the SDH-based joint model (Table 4.9) in contrast to the CSH-based joint model. Individuals with non-cancer AIDS (competing event) were older than those of cancer AIDS (38.6 years vs. 37.2 years). They were censored at the time of failure in the CSH-based joint model, and the square root of their mean CD4+ counts was 13.4. However, they remained in the risk set in the SDH-based joint model, and the square root of their mean CD4+ counts was 18.2. These could be the possible explanation that CD4+ counts increased with the increment of age in the SDH-based joint model for cancer AIDS.

Table 4.9: Joint modeling of longitudinal and survival outcomes (Cancer AIDS)

Joint model	**Using Cox cause-specific hazards (CSH) model**		**Using subdistribution hazards (SDH) model**	
Longitudinal submodel	**Estimate(SE)**	**p-value**	**Estimate(SE)**	**p-value**
Intercept	19.54 (0.25)	<0.0001	16.78 (0.23)	<0.0001
Time	0.73 (0.02)	<0.0001	0.72 (0.03)	<0.0001
MSM	1.09 (0.16)	<0.0001	1.08 (0.14)	<0.0001
Age at diagnosis	-0.07 (0.01)	<0.0001	0.03 (0.01)	<0.0001
Survival submodel	**Estimate(SE)**	**HR (95% CI)**	**Estimate(SE)**	**HR (95% CI)**
MSM	1.51 (0.56)	4.52 (1.51, 13.52)*	1.56 (0.57)	4.76 (1.54, 14.72)*
Age at diagnosis	0.05 (0.01)	1.05 (1.02, 1.08)*	0.05 (0.01)	1.05 (1.02, 1.08)*
Association	-0.21 (0.04)	0.81 (0.75, 0.88)*	-0.20 (0.04)	0.82 (0.75, 0.89)*

*Significant at 5% level of significance

Non-cancer AIDS:

In the joint model of the longitudinal and CSH submodels for non-cancer AIDS, MSM and age were significant in the longitudinal submodel (Table 4.10). MSM and age were not significant in the survival submodel. The association parameter was significant; a unit increase in CD4+ counts corresponded to a 20% decrease in the risk for non-cancer AIDS (HR = 0.80; 95% CI: 0.78-0.82). In the joint model of the longitudinal and SDH submodels, results in both longitudinal and survival submodels were similar but not identical to the results of the CSH-based joint model. Estimates of the MSM were numerically different in the longitudinal submodels. The values of AIC were 48485 and 49438 in the CSH-based and SDH-based joint models, respectively.

Table 4.10: Joint modeling of longitudinal and survival outcomes (Non-Cancer AIDS)

Joint model	**Using Cox cause-specific hazards (CSH) model**		**Using subdistribution hazards (SDH) model**	
Longitudinal submodel	**Estimate(SE)**	**p-value**	**Estimate(SE)**	***P*-value**
Intercept	19.56 (0.25)	<0.0001	19.61 (0.22)	<0.0001
Time	0.72 (0.02)	<0.0001	0.71 (0.02)	<0.0001
MSM	1.10 (0.15)	<0.0001	1.18 (0.11)	<0.0001
Age at diagnosis	-0.07 (0.01)	<0.0001	-0.07 (0.005)	<0.0001
Survival sub model	**Estimate(SE)**	**HR (95% CI)**	**Estimate(SE)**	**HR (95% CI)**
MSM	-0.17 (0.19)	0.84 (0.58, 1.23)	-0.20 (0.19)	0.82 (0.56, 1.19)
Age at diagnosis	0.01 (0.01)	1.01 (1.00, 1.02)	0.01 (0.01)	1.01 (1.00, 1.02)
Association	-0.22 (0.02)	0.80 (0.78, 0.82)*	-0.22 (0.02)	0.80 (0.78, 0.82)*

*Significant at 5% level of significance

Since the association parameter was highly significant in the joint models for cancer AIDS and non-cancer AIDS, this provided strong evidence that both survival outcomes were associated with the longitudinal trajectory of CD4+ counts. Hence, joint modeling was appropriate. In the separate mixed model analyses, age was not significant (Table 4.4). However, age was significant in all longitudinal submodels of joint analyses. Furthermore, in separate analyses, MSM had significantly lower hazards of non-cancer AIDS. Hazards ratios were 0.52 in both multivariable CSH and SDH models (Table 4.6 and Table 4.8). Nevertheless, in joint analyses, MSM had lower hazards of non-cancer AIDS but not significantly lower. Hazards ratios were 0.84 and 0.82 in CSH and SDH submodels, respectively (Table 4.10). Thus, joint analyses provided different results from separate analyses. Also, results between the CSH-based joint model and the SDH-based joint model were consistent for identifying significant covariates but provide slightly different estimates of the covariates for both cancer and non-cancer AIDS.

4.6 Model diagnostics

I discussed joint model diagnostics in Section 3.5.6. I used standardized marginal residuals plots to assess the assumptions of the linear mixed-effects models. In the plot of the standardized marginal residuals versus the fitted values, we noticed that the fitted loess curve (smooth curve between two variables) shows a systematic trend with more negative residuals for large fitted values (Figures 4.7-4.10). Nonetheless, large fitted values corresponded to high levels of square root CD4+ count. Patients with high levels of CD4+ counts, as well as low levels of CD4+ counts, are more likely to be lost to follow-up or dropout (Charurat et al., 2010). From

these figures, we can conclude that the systematic trend could be an indication of the patients' dropout rather than a model lack-of-fit.

Rizopoulos et al. (2010) calculated residuals by imputing missing longitudinal responses under the fitted joint model. However, in the current joint modeling package of R (**JM**), multiple imputation residuals for the longitudinal process cannot be calculated for the Cox survival submodel (Rizopoulos, 2010). For the survival submodel diagnostic, Martingale residuals and Cox-Snell residuals are also not available for Cox submodel in the current **JM** package (Barlow and Prentice, 1988; Cox and Snell, 1968; Rizopoulos, 2010; Therneau et al., 1990). Thus, I was not able to provide Martingale residuals and Cox-Snell residuals plots for the survival submodel diagnostic.

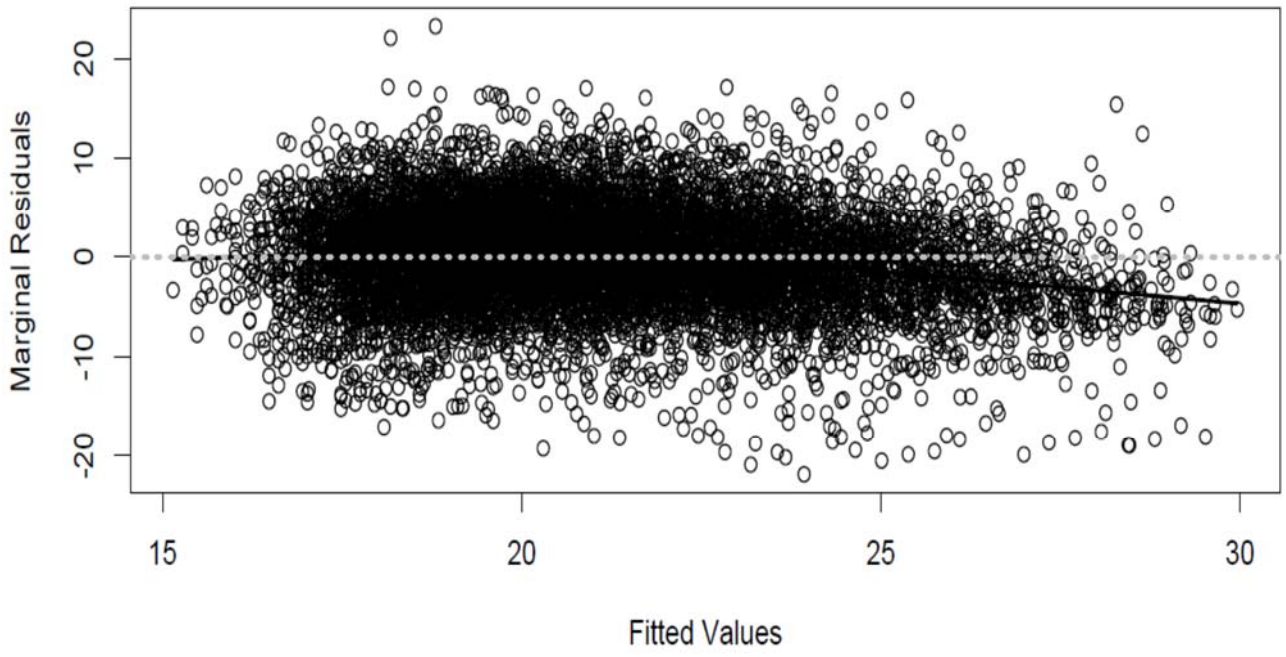


Figure 4.7: Diagnostic plots for the fitted joint model using CSH submodel for Cancer AIDS

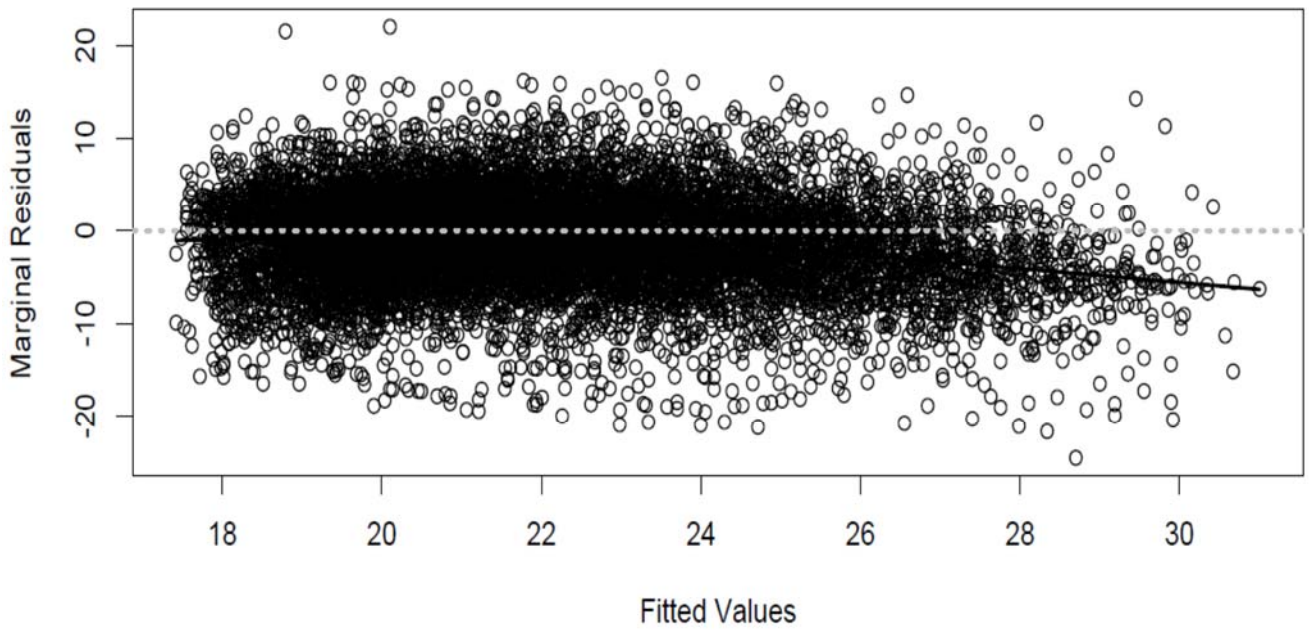


Figure 4.8: Diagnostic plots for the fitted joint model using SDH submodel for Cancer AIDS

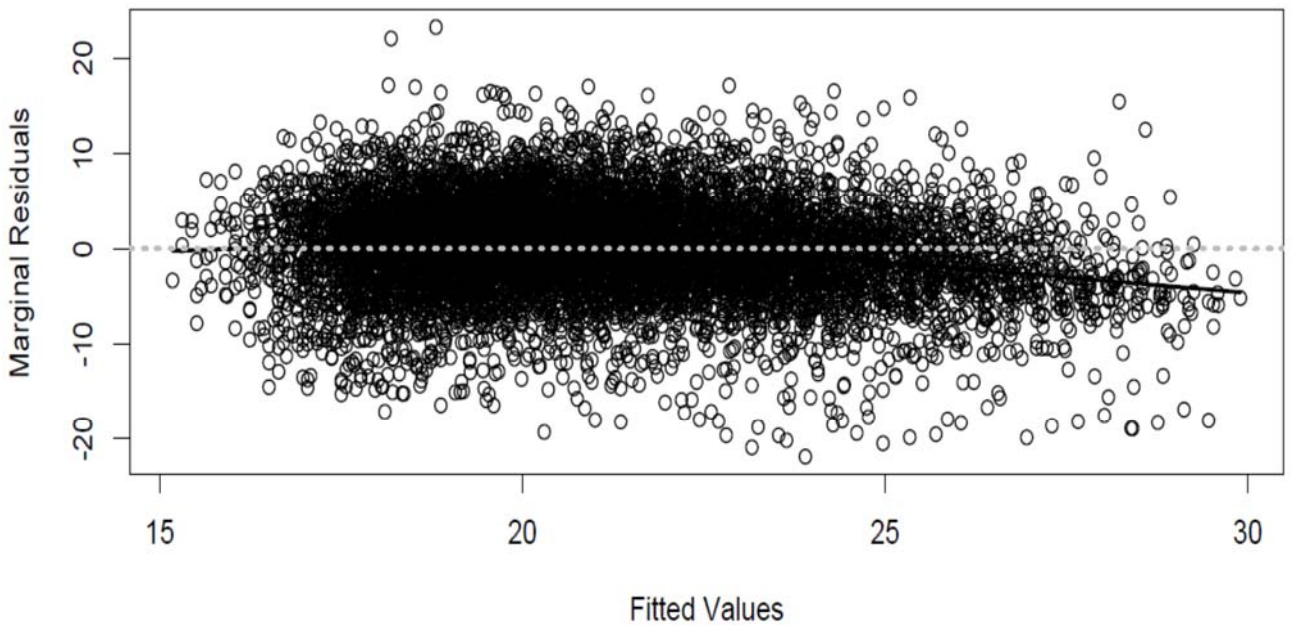


Figure 4.9: Diagnostic plots for the fitted joint model using CSH submodel for non-Cancer AIDS

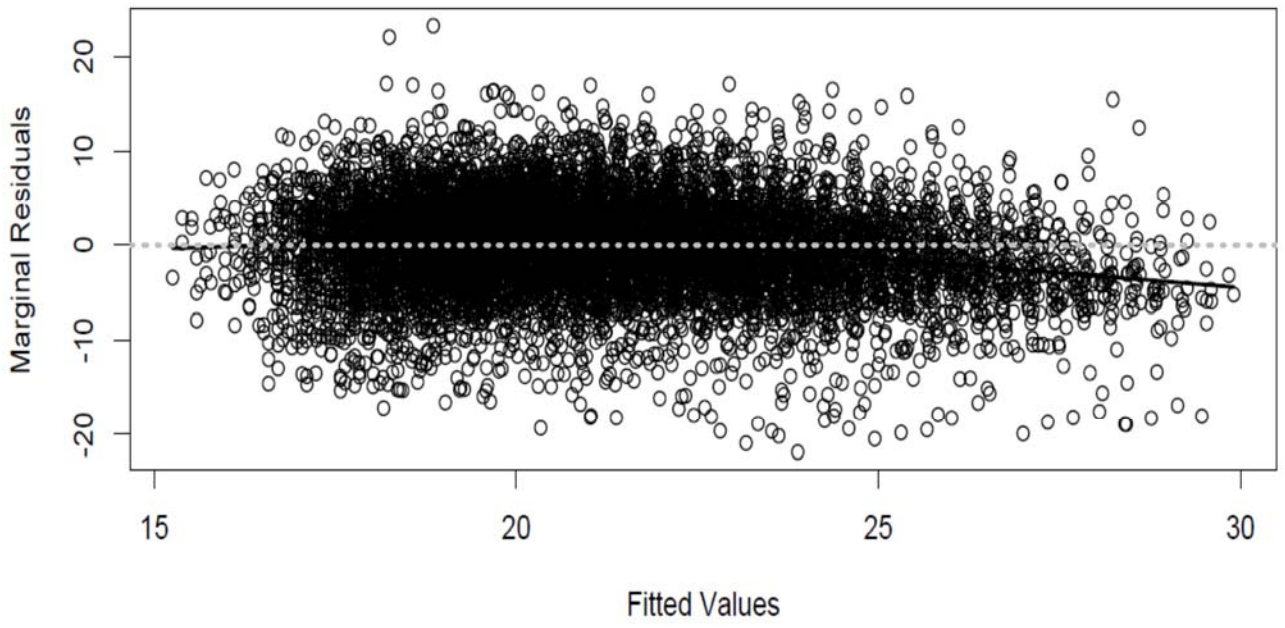


Figure 4.10: Diagnostic plots for the fitted joint model using SDH submodel for non-Cancer AIDS

CHAPTER 5

SIMULATION STUDY

5.1 Simulation study for competing risk

5.1.1 Design of the simulation study for competing risk

I performed this simulation study to accomplish Objective 2, as stated in Chapter 1: To examine the appropriateness of using cause-specific hazards and subdistribution hazards models from the simulation study. The purpose of this simulation was to show how different competing risks scenarios affect results in cause-specific hazards and subdistribution hazards models. The R code for generating competing risks data was obtained from Beyersmann, Allignol, and Schumacher, 2012.

Competing risks data are generated based on the following algorithm from Beyersmann et al. 2009, 2012:

1. The cause-specific hazards $h_{01}(t)$ and $h_{02}(t)$ are assigned as functions of time.
2. Failure times T^* are simulated with all-cause hazard $h_{0.}(t) = h_{01}(t) + h_{02}(t)$.
3. A binomial experiment is performed for a simulated failure time T^*, with probability $h_{0e}(T^*)/[h_{01}(T^*) + [h_{02}(T^*)]$ for cause e, $e = 1, 2$.
4. Right censoring times C are generated.

If a convenience function for the specified cause-specific hazards is not available, inversion method can be used to simulate failure times (Beyersmann et al., 2012). Suppose that $h_{0.}(t) > 0$ for all t, then the all-cause cumulative hazards $H_{0.}(t) = \int_0^t h_{0.}(u)du$ and the distribution

function of T^* are invertible (Bender, Augustin, and Blettner, 2005; Beyersmann et al., 2012). The distribution function of T^* can be expressed as (Beyersmann et al., 2012):

$$F(t) = P(T^* \leq t) = 1 - \exp\big(-H_{0.}(t)\big).$$

Let $F(T^*)$ be the transformed failure time. Then based on the principle of the inversion method, $F(T^*)$ is uniformly distributed on [0, 1] (Bender et al., 2005; Beyersmann et al., 2012):

$$P(F(T^*) \leq u) = P\big(T^* \leq F^{-1}(u)\big) = F\big(F^{-1}(u)\big) = u, u \in [0,1]$$

Therefore, if the random variable U has uniform distribution on [0, 1], then the distribution of $F^{-1}(U)$ will be same as T^* (Beyersmann et al., 2012). The following steps are followed in the application of inversion method (Beyersmann et al., 2012):

1. We calculate $F^{-1}(u) = H_{0.}^{-1}(-\ln(1-\mathrm{u})),\ \mathrm{u} \in [0,1]$.
2. A random variable U with uniform distribution on [0, 1] is generated
3. $F^{-1}(U)$ gives the simulated failure times of T^*.

I shall fit CSH and SDH models using simulated data and explore if the results between CSH and SDH models are different for different proportions of the main event of interest and competing risks.

5.1.2 Model specification

Competing risks data were simulated based on cause-specific hazards, where the hazards were allowed to be time-dependent (Beyersmann et al., 2009). Data were simulated by roughly mimicking the HIV study data from the OCS presented in Chapter 4. Cause-specific hazard plots

of cancer AIDS and non-cancer AIDS of the study data are presented in Figure 5.1 and Figure 5.2, respectively, by HIV risk category. Based on cause-specific hazards in Figure 5.1 and Figure 5.2, I roughly imitated the following fractional polynomials (Royston and Altman, 1994) for the cause-specific baseline hazards of cancer AIDS and non-cancer AIDS, respectively:

$$h_{i1;0}(t) = h_{i1;0}(t; a = 0) = \frac{0.008}{t+1}, \tag{5.1}$$

$$h_{i2;0}(t) = h_{i2;0}(t; a = 0) = \frac{0.035}{t+1}, \qquad i = 1, 2, \dots, n. \tag{5.2}$$

where a indicates a baseline covariate, such as the HIV risk category indicator, with values 0 and 1. I assume the following separate proportional CSH models for the effect of $a = 1$:

$$h_{i1}(t; a = 1) = 3.065 \cdot h_{i1;0}(t), \tag{5.3}$$

$$h_{i2}(t; a = 1) = 0.533 \cdot h_{i2;0}(t), i = 1, 2, \dots, n. \tag{5.4}$$

I chose the models (5.3) and (5.4) based on the study data, where $a = 1$ indicates MSM and $a = 0$ indicates all other individuals. From the study data, for MSM compared to the other HIV risk category, I obtained hazards ratios of cancer AIDS and non-cancer AIDS of 3.065 and 0.533, respectively. There were 822 individuals in our study data. Among them, 22 (2.7%) had cancer AIDS, 123 (15.0%) had non-cancer AIDS, and 677 (82.4%) were censored. Four hundred eighty-one (58.5%) were MSM.

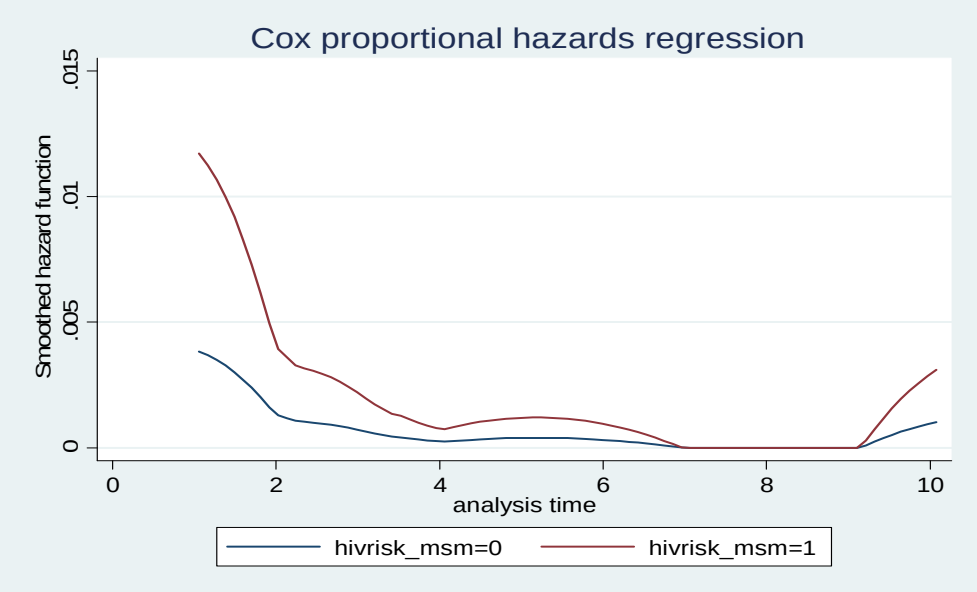


Figure 5.1: Cause-specific hazards plot for cancer AIDS by HIV risk group: MSM and Other

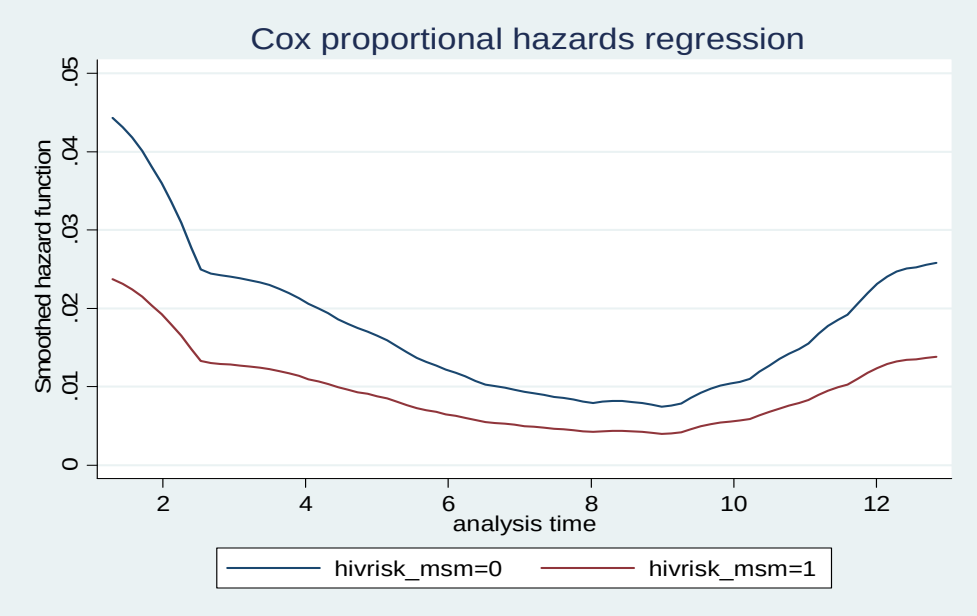


Figure 5.2: Cause-specific hazards plot for non-cancer AIDS by HIV risk group: MSM and Other

I simulated 825 individuals with 495 (60%) as MSM by roughly mimicking the OCS study data. However, like the study data, I was not able to simulate data with similar proportions of

cancer AIDS (2.7%) and non-cancer AIDS (15.0%). Estimates were very unstable when I simulated data with a much lower number of events for cancer AIDS. Table 5.1 showed 20 individuals who were selected randomly from the simulated data.

Table 5.1: Randomly selected 20 individuals from the competing risks simulated data

ID	HIV risk Category[§]	Survival time (years)	Status[£]	ID	HIV risk category	Survival time (years)	Status
662	1	2.1	2	466	1	9.4	0
245	0	5.7	2	526	1	0.2	1
360	1	5.1	0	445	1	2.8	0
540	1	2.7	0	740	1	5.7	0
503	1	0.1	1	156	0	15.4	0
97	0	0.5	1	527	1	7.9	0
460	1	6.0	0	249	0	9.1	0
303	0	1.4	0	201	0	6.5	0
563	1	10.2	0	289	0	1.2	2
646	1	9.7	0	681	1	2.2	0

[§]1 indicates MSM and 0 indicates other; [£]1 indicates cancer-AIDS, 2 indicates non-cancer AIDS, and 0 indicates censored;

In addition to the cause-specific baseline hazards defined in (5.1) and (5.2), I considered two more scenarios to get different proportions of events and competing events. I conducted 500, 1000 replications for each of the three scenarios and fitted both CSH and SDH models. For a specific seed (seed # = 261139), the mean proportions of individuals with cancer AIDS, non-

cancer AIDS, and censored were *14.6%, 20.6%, and 64.8%,* respectively, in scenario 1; *15.8%, 25.0%, and 59.2%,* respectively, in scenario 2; and *16.1%, 29.1%, and 54.8%,* respectively, in scenario 3 (Table 5.2 and 5.3).

Table 5.2 shows the mean parameter estimates, bias, mean standard errors, and the 95% confidence interval coverage probabilities of the estimates for MSM when the event is cancer AIDS, and the competing event is non-cancer AIDS. The mean parameter estimates of 1.13, 1.18, and 1.20 in scenarios 1, 2, and 3, respectively, are reasonably close to the true value 1.12. Furthermore, the 95% confidence interval coverage probabilities are about 95% in all scenarios. Estimated parameters in SDH models are numerically different from CSH models in all scenarios.

In the analysis of real data, the standard error of the estimate of MSM effect on cancer AIDS was higher (0.55) in the CSH model (Table 4.5). As a result, the 95% confidence interval of the estimate of MSM was wider. However, in the analysis of simulated data, the value of the corresponding standard error is much smaller in all scenarios (Table 5.2).

Table 5.2: Comparison of estimates for MSM between CSH and SDH models (event = cancer, competing event = non-cancer)

				CSH				SDH	
Scenario	**Number of replication**	**Mean number of cancer event**	**Mean number of non-cancer event**	**Mean estimate[a]**	**Bias**	**Mean SE[¥]**	**95% CP[§]**	**Mean estimate**	**Mean SE**
Scenario 1[b]	500	120 (14.6%)	169 (20.5%)	1.13457	0.01452	0.24105	0.972	1.22005	0.24109
	1000	120 (14.6%)	170 (20.6%)	1.13069	0.01064	0.24089	0.965	1.21498	0.24092
Scenario 2[c]	500	131 (15.9%)	206 (25.0%)	1.18576	0.06571	0.23573	0.940	1.28469	0.23577
	1000	130 (15.8%)	206 (25.0%)	1.18030	0.06025	0.23561	0.949	1.27846	0.23565
Scenario 3[d]	500	132 (16.0%)	242 (29.3%)	1.20554	0.08549	0.23698	0.940	1.31526	0.23701
	1000	133 (16.1%)	240 (29.1%)	1.20410	0.08405	0.23643	0.943	1.31448	0.23646

[a]True value, log (3.065) = 1.12005 in CSH model for cancer; [¥]Standard Error; [§]Coverage Probability

[b]Cause-specific baseline hazard function: $h_{i1;0}(t) = h_{i1;0}(t; a = 0) = \frac{0.008}{t+1}$, $h_{i2;0}(t) = h_{i2;0}(t; a = 0) = \frac{0.035}{t+1}$

[c]Cause-specific baseline hazard function: $h_{i1;0}(t) = h_{i1;0}(t; a = 0) = \frac{0.022}{t+1}$, $h_{i2;0}(t) = h_{i2;0}(t; a = 0) = \frac{0.110}{t+1}$

[d]Cause-specific baseline hazard function: $h_{i1;0}(t) = h_{i1;0}(t; a = 0) = \frac{0.030}{t+1}$, $h_{i2;0}(t) = h_{i2;0}(t; a = 0) = \frac{0.175}{t+1}$

Mean parameter estimates, bias, mean standard errors, and the 95% CI coverage probabilities of the estimates for MSM, when the event is non-cancer AIDS, and the competing event is cancer AIDS, are reported in Table 5.3. We noticed that mean parameter estimates -0.63, -0.60, -0.57 in scenarios 1, 2, and 3, respectively, are reasonably close to the true value of -0.63. Also, the 95% CI coverage probabilities are about 95% in all scenarios. Like Table 5.2, estimated parameters in SDH models are numerically different from CSH models in all scenarios.

Notably, I simulated data based on CSH models. Thus, I compared estimates from CSH models with the true value and reported bias and the 95% confidence interval coverage

probabilities. If the proportional hazards assumption is met in CSH model, this assumption does not hold in SDH model (Beyersmann et al., 2009; Latouche et al., 2013). I fitted SDH models with the censoring complete data that were simulated based on CSH models. Hence, bias and the 95% confidence interval coverage probabilities are not reported for SDH models in Table 5.2 and Table 5.3. Here my intention is to show if SDH models provide different estimates from CSH models in different competing risks scenarios.

In addition, I could simulate competing risks data based on SDH models and compare estimates between SDH and CSH models. However, this is one of my future research interests.

Table 5.3: Comparison of estimates between CSH and SDH models (event = non-cancer, competing event = cancer)

				CSH				**SDH**	
Scenario	**Number of replication**	**Mean number of cancer event**	**Mean number of non-cancer event**	**Mean estimate[a]**	**Bias**	**Mean SE[¥]**	**95% CP[§]**	**Mean estimate**	**Mean SE**
Scenario 1[b]	500	120 (14.6%)	169 (20.5%)	-0.64440	0.01517	0.15549	0.956	-0.73665	0.15556
	1000	120 (14.6%)	170 (20.6%)	-0.63229	0.00306	0.15522	0.957	-0.72328	0.15529
Scenario 2[c]	500	131 (15.9%)	206 (25.0%)	-0.59957	0.02966	0.14045	0.954	-0.70509	0.14054
	1000	130 (15.8%)	206 (25.0%)	-0.59881	0.03042	0.14026	0.958	-0.70340	0.14035
Scenario 3[d]	500	132 (16.0%)	242 (29.3%)	-0.56709	0.06214	0.12944	0.940	-0.67782	0.12955
	1000	133 (16.1%)	240 (29.1%)	-0.57385	0.05538	0.12979	0.944	-0.68465	0.12990

[a]True value, log (0.533) = -0.62923 in CSH model for non-cancer; [¥]Standard Error; [§]Coverage Probability

[b]Cause-specific baseline hazard function: $h_{i1;0}(t) = h_{i1;0}(t; a = 0) = \frac{0.008}{t+1}$, $h_{i2;0}(t) = h_{i2;0}(t; a = 0) = \frac{0.035}{t+1}$

[c]Cause-specific baseline hazard function: $h_{i1;0}(t) = h_{i1;0}(t; a = 0) = \frac{0.022}{t+1}$, $h_{i2;0}(t) = h_{i2;0}(t; a = 0) = \frac{0.110}{t+1}$

[d]Cause-specific baseline hazard function: $h_{i1;0}(t) = h_{i1;0}(t; a = 0) = \frac{0.030}{t+1}$, $h_{i2;0}(t) = h_{i2;0}(t; a = 0) = \frac{0.175}{t+1}$

Summary of the simulations:

I studied three scenarios using three different cause-specific baseline hazards for both cancer AIDS and non-cancer AIDS. Although I simulated data roughly mimicking the real OCS HIV data, the proportions of both events in the simulated data were different from the real data.

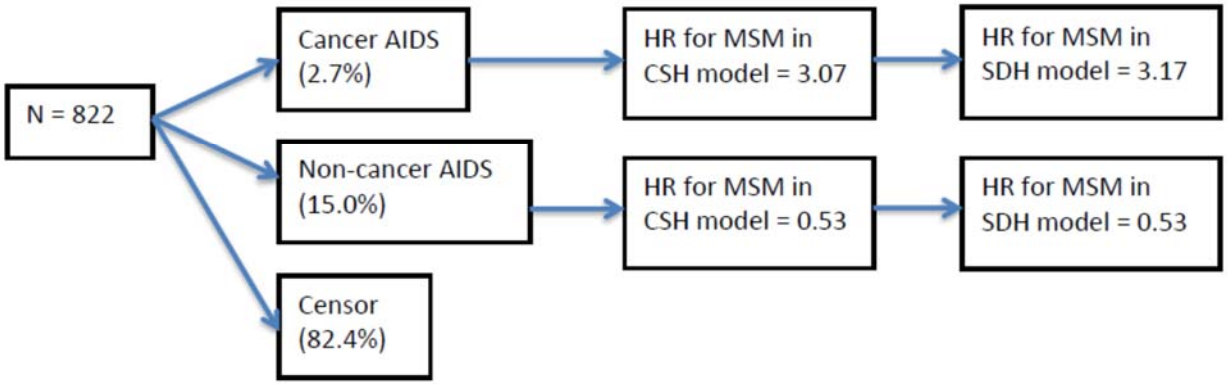
In the real HIV data, the proportion of cancer AIDS was very low (2.7%), while the proportion of non-cancer AIDS was relatively very highly (15.0%). The proportions of both events were higher in the simulated data compared to the real data. Proportions of cancer AIDS were 14.6%, 15.8%, and 16.1% in scenario 1, scenario 2, and scenario 3, respectively. Proportions of non-cancer AIDS were 20.6%, 25.0%, and 29.1% in scenario 1, scenario 2, and scenario 3, respectively.

In the real data, hazards ratios of cancer AIDS for MSM were numerically different between the CSH and SDH models (3.07 and 3.17, respectively; Figure 5.3, i). However, this difference was bigger in the simulated data. Hazards ratios in the CSH and SDH models were 3.10 and 3.37, respectively, in Scenario 1 (Figure 5.3, ii); 3.26 and 3.59, respectively, in Scenario 2 (Figure 5.3, iii); 3.33 and 3.72, respectively, in Scenario 3 (Figure 5.3, iv).

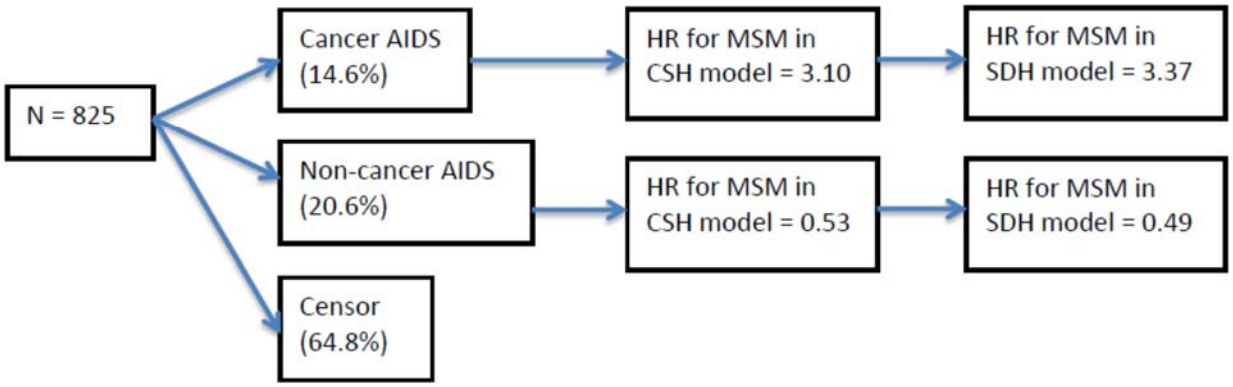
In the real data, the proportion of the competing event (cancer AIDS) was very low in the SDH model of non-cancer AIDS. Hence, cancer AIDS as a competing event did not influence the subdistribution hazards model of non-cancer AIDS. Hazards ratios of non-cancer AIDS for MSM were equal (0.53) in CSH and SDH models. On the contrary, hazards ratios of non-cancer AIDS for MSM were numerically different in the CSH and SDH models in the simulated data. Hazards ratios in the CSH and SDH models were 0.53 and 0.49, respectively, in Scenario 1 (Figure 5.3, ii);

0.55 and 0.49, respectively, in Scenario 2 (Figure 5.3, iii); 0.56 and 0.50, respectively, in Scenario 3 (Figure 5.3, iv). Thus, results from the simulation study indicated that SDH model would provide different results than the CSH model even if the proportion of the competing event is not very much lower than the event of interest.

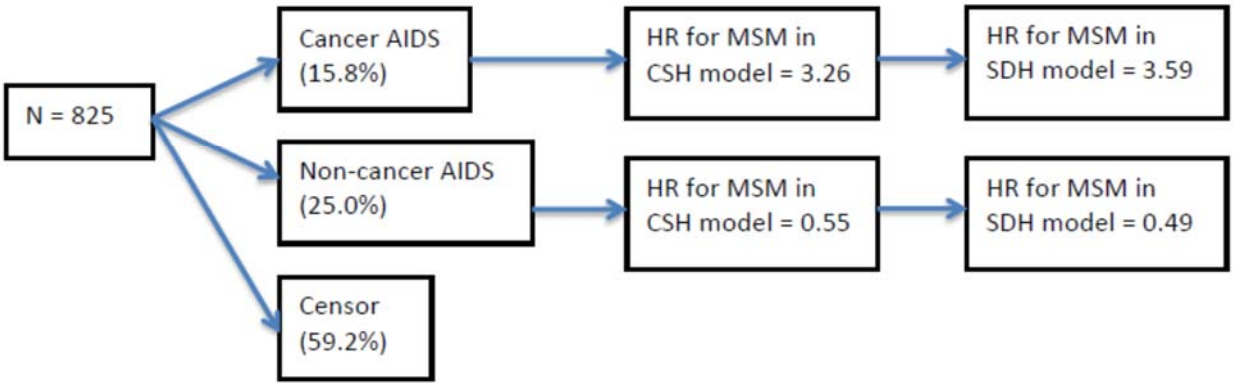
(i) Results from real HIV data (N = 822)



(ii) Simulated data (scenario 1): true HRs for cancer and non-cancer are 3.07 and 0.53, respectively in CSH models.



(iii) Simulated data (scenario 2): true HRs for cancer and non-cancer are 3.07 and 0.53, respectively in CSH models.



(iv) Simulated data (scenario 3): true HRs for cancer and non-cancer are 3.07 and 0.53, respectively in CSH models.

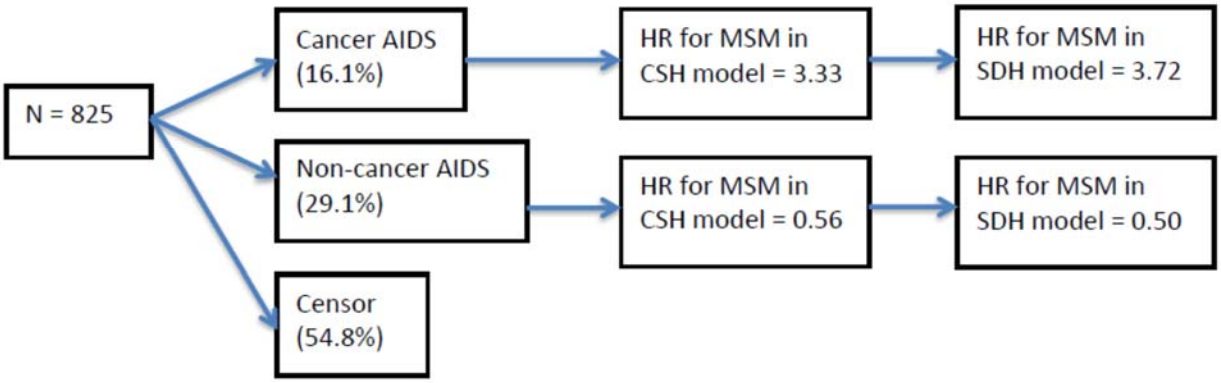


Figure 5.3: Comparison of hazards ratios between CSH and SDH models for real data and simulated data
Note: All scenarios are based on 1000 replications.

5.2 Joint model simulation

This simulation study was performed to accomplish Objective 3. The purpose of this simulation was to investigate how magnitudes of association between the longitudinal marker and the time-to-event outcome influence parameters estimate in separate Cox regression

models and linear mixed effects models. The SAS macro for joint model simulation was obtained from Ibrahim et al. (2010).

5.2.1 Simulation design

The simulation was conducted based on the joint model discussed in Chapter 3. The simulation study was performed with $\gamma = 0.5$ and $\beta = 0.5$, where γ and β indicate the treatment effect on survival and longitudinal outcomes, respectively (Ibrahim et al., 2010). Several magnitudes of association parameter $\alpha = 0.01, 0.1, 0.2, 0.3, 0.4, 0.5$ between longitudinal and survival outcomes were considered to see how these values influence the estimates of the parameters in separate Cox and linear mixed effects models. With the above parameters, I performed 500 replications. In each replication, I generated 200, 400, and 500 individuals, with an equal number of individuals each in the treatment and control groups. I simulated true longitudinal trajectory m_{ij} as (Ibrahim et al., 2010):

$$m_{ij} = \beta_0 + \beta_1 \times (time)_{ij} + \beta_2 \times treat_i + b_{0i} + b_{1i} \times (time)_{ij},$$

$$i = 1,2 \dots, n,\ j = 1,2 \dots, m_i,$$

where $b_{0i} \sim N(0, 1)$, $b_{1i} \sim N(0, 0.6)$, and the correlation between b_{0i} and b_{1i} was chosen as 0.15. I simulated the observed longitudinal data from the model $y_i(t_j) = N(m_i(t_j),\ \sigma_\varepsilon^2)$ with $\sigma_\varepsilon^2 = 0.6^2$ (Ibrahim et al., 2010). A maximum of six measurements at times $t_j = (0, 0.5, 1.0, 1.5, 2.0, 2.5)$ years was considered for the longitudinal data. Similar to Ibrahim et al. (2010), I used inverse probability method to generate survival time with the assumption of

uniform right censoring in the interval [1, 3]. I considered a constant baseline hazard of 0.30 to generate survival time from the equation (Ibrahim et al., 2010):

$$h_i(t) = h_0(t)\exp\{\gamma' a_i + \alpha m_{ij}\}.$$

The choice of baseline hazard depends on the number of events to be generated. For example, Sweeting and Thompson (2011) wanted to simulate a rare event in their study. Consequently, they chose the value for constant baseline hazard as 0.008. Table 5.4 showed 10 subjects that were selected randomly from the simulated data where sample size and association parameter were 200 and 0.20, respectively.

Table 5.4: Randomly selected 10 subjects from the simulated data

ID	Time	Longitudinal Data	Treatment[¥]	Survival time	Status[§]	ID	Time	Longitudinal data	Treatment	Survival time	Status
9	0	-1.54	0	1.64	0	94	0	-0.41	0	2.63	1
9	0.5	0.63	0	1.64	0	94	0.5	0.77	0	2.63	1
9	1.0	1.39	0	1.64	0	94	1.0	2.32	0	2.63	1
9	1.5	4.60	0	1.64	0	94	1.5	4.29	0	2.63	1
28	0	-0.47	0	2.18	1	94	2.0	5.46	0	2.63	1
28	0.5	0.35	0	2.18	1	94	2.5	7.42	0	2.63	1
28	1.0	1.89	0	2.18	1	122	0	-2.04	1	1.0	0
28	1.5	3.38	0	2.18	1	122	0.5	-2.31	1	1.0	0
28	2.0	4.38	0	2.18	1	122	1.0	-0.03	1	1.0	0
37	0	-0.01	0	2.0	0	175	0	-0.26	1	0.45	1
37	0.5	1.17	0	2.0	0	196	0	-0.04	0	2.25	0
37	1.0	2.22	0	2.0	0	196	0.5	1.48	0	2.25	0
37	1.5	4.58	0	2.0	0	196	1.0	3.16	0	2.25	0
59	0	0.33	0	1.0	0	196	1.5	2.61	0	2.25	0

59	0.5	1.42	0	1.0	0	196	2.0	3.73	0	2.25	0
59	1.0	2.15	0	1.0	0	198	0	-0.07	0	1.78	1
61	0	-0.15	1	1.69	1	198	0.5	1.97	0	1.78	1
61	0.5	2.00	1	1.69	1	198	1.0	4.06	0	1.78	1

[¥]0 indicates no and 1 indicates yes; [§]0 indicates censored, and 1 indicates an event.

Three models were fitted for each simulation: (1) a simple Cox model; (2) a linear mixed effects model; (3) the joint model. A comparison of the estimates of the treatment effect on survival (γ) obtained from separate Cox models and joint models for different magnitudes of association between longitudinal and survival outcomes are shown in Table 5.5.

Table 5.5: Comparison of the estimation of the treatment effect on survival (γ) between separate Cox models and joint models

Sample size	**True parameter value**			**Cox PH model**				**Joint model**			
n	β	γ	α	Mean Estimate ($\hat{\gamma}$)	Mean SE[¥]	Bias	95% CP[§]	Mean Estimate ($\hat{\gamma}$)	Mean SE[¥]	Bias	95% CP[§]
200	**0.50**	**0.50**	**0.01**	0.50791	0.2122	0.0079	95.6	0.5052	0.2120	0.0052	95.0
	0.50	**0.50**	**0.10**	0.55057	0.19544	0.05058	93.4	0.5054	0.1943	0.00536	95.8
	0.50	**0.50**	**0.20**	0.57954	0.18170	0.07955	92.2	0.5000	0.1801	0.00005	93.4
	0.50	**0.50**	**0.30**	0.59684	0.17243	0.09684	91.4	0.4958	0.1711	-0.00419	94.0
	0.50	**0.50**	**0.40**	0.60883	0.16616	0.10883	90.0	0.4985	0.1660	-0.00147	93.4
	0.50	**0.50**	**0.50**	0.61433	0.16155	0.11433	89.0	0.4975	0.1634	-0.00245	94.8
400	**0.50**	**0.50**	**0.01**	0.50367	0.14919	0.00367	95.4	0.4997	0.1491	-0.0003	95.4
	0.50	**0.50**	**0.10**	0.54631	0.13742	0.04632	95.8	0.5009	0.1368	0.00085	96.0
	0.50	**0.50**	**0.20**	0.57680	0.12774	0.07680	91.8	0.4985	0.1268	-0.00145	95.4
	0.50	**0.50**	**0.30**	0.59536	0.12125	0.09536	88.4	0.4964	0.1205	-0.00355	95.6
	0.50	**0.50**	**0.40**	0.60615	0.11681	0.10615	85.6	0.4966	0.1169	-0.00343	95.4
	0.50	**0.50**	**0.50**	0.61221	0.11357	0.11221	84.0	0.5004	0.1152	0.00036	95.8

500	**0.50**	**0.50**	**0.01**	0.50265	0.13331	0.00265	94.0	0.4984	0.1333	-0.00163	94.2
	0.50	**0.50**	**0.10**	0.54536	0.12278	0.04536	94.6	0.4999	0.1223	-0.00008	94.6
	0.50	**0.50**	**0.20**	0.57554	0.11408	0.07554	89.8	0.4981	0.1133	-0.00192	95.8
	0.50	**0.50**	**0.30**	0.59288	0.10825	0.09288	87.2	0.4972	0.1077	-0.00282	95.8
	0.50	**0.50**	**0.40**	0.60391	0.10431	0.10391	82.6	0.4971	0.1045	-0.00292	96.2
	0.50	**0.50**	**0.50**	0.60991	0.10141	0.10991	80.4	0.4985	0.1029	-0.00155	94.0

¥Standard Error; §Coverage Probability
β = Treatment effect on longitudinal outcome; γ = Treatment effect on survival outcome;
α = Association parameter between longitudinal and survival outcomes;
Note: The values of β (0.5) and γ (0.5) are fixed. However, the values of association parameter α are 0.01, 0.1, 0.2, 0.3, 0.4, and 0.5.

Summary from Table 5.5

For all sample sizes and all values of the association parameter, mean estimates of the treatment effect on survival outcome in the joint model are almost equal to the true value (0.50). In all circumstances, the 95% confidence interval coverage probabilities of the estimates are about 95%. When there is no association between longitudinal and survival outcomes (magnitude of association parameter = 0.01), mean estimates of the treatment effect on survival in separate Cox models are also close to the true value (0.50), i.e., the bias of the estimate is minimal. When the magnitude of the association increases, mean estimates of the treatment effect on survival outcome in a joint model do not change; however, mean estimates of the treatment effect in separate Cox models change and bias increases. The bias of the estimate in separate Cox models increases as the magnitude of the association between longitudinal outcome and survival outcome increases. This indicates that when longitudinal and survival outcomes are associated, separate Cox regression analyses provide the bias estimate. The magnitude of the bias depends on the magnitude of the association between longitudinal and survival outcomes.

Notably, mean estimates of the association parameter in the joint model were almost equal to the true values in all circumstances. For sample size 400, and for association parameters 0.20, 0.30, mean estimates of the association parameter in joint model were 0.2032, 0.3039, respectively (95% confidence interval coverage probabilities were 95.0% and 94.8%, respectively).

Table 5.6: Comparison of the estimation of the treatment effect on longitudinal outcome (β) between separate mixed effects models and joint models

Sample size	**True parameter value**			**Mixed effects model**				**Joint model**			
n	β	γ	α	Mean Estimate ($\hat{\beta}$)	Mean SE¥	Bias	95% CP§	Mean Estimate ($\hat{\beta}$)	Mean SE¥	Bias	95% CP§
200	**0.50**	**0.50**	**0.01**	0.5112	0.1593	0.01117	96.2	0.5112	0.1593	0.01119	96.2
	0.50	**0.50**	**0.10**	0.5015	0.1596	0.00146	95.8	0.5024	0.1596	0.00241	96.0
	0.50	**0.50**	**0.20**	0.5070	0.1598	0.00705	96.8	0.5100	0.1598	0.01005	96.4
	0.50	**0.50**	**0.30**	0.5005	0.1602	0.00046	95.8	0.5065	0.1601	0.0065	96.0
	0.50	**0.50**	**0.40**	0.4934	0.1604	-0.00663	96.0	0.5028	0.1603	0.00280	95.6
	0.50	**0.50**	**0.50**	0.4916	0.1607	-0.00838	96.0	0.5046	0.1604	0.00455	95.2
400	**0.50**	**0.50**	**0.01**	0.5070	0.1125	0.00698	95.4	0.5027	0.1126	0.00270	95.2
	0.50	**0.50**	**0.10**	0.5034	0.1128	0.003407	94.2	0.5044	0.1127	0.004417	94.2
	0.50	**0.50**	**0.20**	0.4970	0.1129	-0.00304	95.4	0.5000	0.1129	0.00004	95.6
	0.50	**0.50**	**0.30**	0.4973	0.1132	-0.00274	95.2	0.5034	0.1132	0.00342	95.2
	0.50	**0.50**	**0.40**	0.4979	0.1133	-0.00209	95.6	0.5074	0.1132	0.00735	95.2
	0.50	**0.50**	**0.50**	0.4875	0.1135	-0.01245	96.0	0.5006	0.1133	0.00055	96.4
500	**0.50**	**0.50**	**0.01**	0.4996	0.1007	-0.00041	95.2	0.4997	0.1007	-0.00035	95.2
	0.50	**0.50**	**0.10**	0.5000	0.1008	0.00002	94.6	0.5011	0.1008	0.00107	94.8
	0.50	**0.50**	**0.20**	0.4966	0.1010	-0.00341	94.0	0.4997	0.1010	-0.00035	94.4

	0.50	**0.50**	**0.30**	0.4938	0.1012	-0.00621	94.6	0.4999	0.1011	-0.00013	94.6
	0.50	**0.50**	**0.40**	0.4903	0.1014	-0.00974	94.6	0.4997	0.1013	-0.00025	94.8
	0.50	**0.50**	**0.50**	0.4883	0.1016	-0.01174	94.6	0.5013	0.1014	0.00126	94.4

¥Standard Error, §Coverage Probability
Note: The values of β (0.5) and γ (0.5) are fixed. However, the values of association parameter α are 0.01, 0.10, 0.20, 0.30, 0.40, and 0.50.

Summary from Table 5.6

A comparison of the estimate of the treatment effect on the longitudinal outcome (β) obtained from separate mixed effects models and joint models for different magnitudes of association are shown in Table 5.6. Mean estimates of the treatment effect on longitudinal outcome in joint models are almost equal to the true value (0.50) for all sample sizes and all association parameters. In all situations, the 95% confidence interval coverage probabilities of the estimate are about 95%. Thus, estimates are unbiased in the joint model. Mean estimates of the treatment effect are almost similar in joint models and separate mixed effects models. As a side note, I should mention that I simulated data based on a joint model similar to Wulfsohn and Tsiatis (1997), Rizopoulos (2010, 2012), and Ibrahim et al. (2010). In the joint model in this setting, the main objective is to accomplish inference for the survival outcome while considering the impact of time-dependent repeatedly measured outcomes measured with error.

Negative association:

I also performed simulation studies on the joint model using negative association between longitudinal and survival outcomes (Table 5.7, Table 5.8). I chose several magnitudes of association parameter $\alpha =$ -0.01, -0.1, -0.2, -0.3, -0.4, -0.5 between the two outcomes. Results were consistent with the simulation studies using positive association in Table 5.7 and Table

5.8. When the association between longitudinal and survival outcomes increases, mean estimates of the treatment effect on survival in separate Cox models change and bias increases (Table 5.7). However, the bias of the estimate in separate Cox models was larger when I used negative association between longitudinal and survival outcomes. Mean estimates of the treatment effect on longitudinal outcome looked similar in joint model and separate mixed effects model (Table 5.8).

Table 5.7: Comparison of the estimation of the treatment effect on survival (γ) between separate Cox models and joint models (negative association)

Sample size	**True parameter value**			**Cox PH model**				**Joint model**			
n	β	γ	α	Mean Estimate ($\hat{\gamma}$)	Mean SE¥	Bias	95% CP§	Mean Estimate ($\hat{\gamma}$)	Mean SE¥	Bias	95% CP§
200	**0.50**	**0.50**	**-0.01**	0.49731	0.21622	-0.00269	95.8	0.5063	0.2179	0.00629	95.8
	0.50	**0.50**	**-0.10**	0.45267	0.23282	-0.04733	95.0	0.5101	0.2343	0.01008	93.8
	0.50	**0.50**	**-0.20**	0.39001	0.24691	-0.10999	94.8	0.5075	0.2488	0.00754	95.0
	0.50	**0.50**	**-0.30**	0.32197	0.25502	-0.17803	91.0	0.5025	0.2578	0.00253	94.4
	0.50	**0.50**	**-0.40**	0.25469	0.25883	-0.24531	85.6	0.5044	0.2634	0.00437	95.0
	0.50	**0.50**	**-0.50**	0.19388	0.26023	-0.30612	77.8	0.5110	0.2675	0.01096	93.0
400	**0.50**	**0.50**	**-0.01**	0.49190	0.15197	-0.00810	95.2	0.4981	0.1531	-0.00190	95.4
	0.50	**0.50**	**-0.10**	0.44549	0.16354	-0.05452	93.6	0.5002	0.1647	0.00024	95.6
	0.50	**0.50**	**-0.20**	0.38544	0.17325	-0.11456	90.4	0.4986	0.1747	-0.00143	95.2
	0.50	**0.50**	**-0.30**	0.32046	0.17899	-0.17954	83.6	0.4966	0.1811	-0.00336	94.2
	0.50	**0.50**	**-0.40**	0.25812	0.18165	-0.24188	74.6	0.4980	0.1849	-0.00198	94.6
	0.50	**0.50**	**-0.50**	0.19932	0.18238	-0.30068	62.4	0.5008	0.1876	0.00082	94.8
500	**0.50**	**0.50**	**-0.01**	0.49143	0.13581	-0.00857	93.6	0.4974	0.1369	-0.00263	94.0
	0.50	**0.50**	**-0.10**	0.44619	0.14613	-0.05381	92.6	0.4999	0.1471	-0.00007	95.0
	0.50	**0.50**	**-0.20**	0.38409	0.15478	-0.11591	88.0	0.4968	0.1560	-0.00320	95.8
	0.50	**0.50**	**-0.30**	0.31930	0.15988	-0.18070	79.0	0.4950	0.1618	-0.00498	95.2

	0.50	**0.50**	**-0.40**	0.25678	0.16230	-0.24322	66.4	0.4965	0.1652	-0.00347	95.0
	0.50	**0.50**	**-0.50**	0.19901	0.16293	-0.30099	56.8	0.4984	0.1675	-0.00160	94.2

¥Standard Error; §Coverage Probability

β = Treatment effect on longitudinal outcome; γ = Treatment effect on survival outcome;

α = Association parameter between longitudinal and survival outcomes;

Note: The values of β (0.5) and γ (0.5) are fixed. However, the values of association parameter α are -0.01, -0.1, -0.2, -0.3, -0.4, and -0.5.

Table 5.8: Comparison of the estimation of the treatment effect on longitudinal outcome (β) between separate mixed effects models and joint models (negative association)

Sample size	**True parameter value**			**Mixed effects model**				**Joint model**			
n	β	γ	α	Mean Estimate ($\hat{\beta}$)	Mean SE¥	Bias	95% CP§	Mean Estimate ($\hat{\beta}$)	Mean SE¥	Bias	95% CP§
200	**0.50**	**0.50**	**-0.01**	0.5053	0.1593	0.00533	96.0	0.5053	0.1593	0.00531	96.0
	0.50	**0.50**	**-0.10**	0.5079	0.1589	0.00792	96.2	0.5077	0.1589	0.00771	96.2
	0.50	**0.50**	**-0.20**	0.5081	0.1585	0.00811	96.0	0.5080	0.1585	0.00801	96.0
	0.50	**0.50**	**-0.30**	0.5099	0.1583	0.00990	96.2	0.5100	0.1582	0.01001	96.0
	0.50	**0.50**	**-0.40**	0.5121	0.1582	0.01208	95.4	0.5123	0.1582	0.01231	95.4
	0.50	**0.50**	**-0.50**	0.5097	0.1582	0.00966	96.8	0.5099	0.1582	0.00994	97.0
400	**0.50**	**0.50**	**-0.01**	0.5032	0.1125	0.00325	95.6	0.5032	0.1125	0.00320	95.6
	0.50	**0.50**	**-0.10**	0.5030	0.1122	0.00297	94.4	0.5027	0.1122	0.00266	94.4
	0.50	**0.50**	**-0.20**	0.5075	0.1120	0.00750	94.5	0.5072	0.1119	0.00722	94.4
	0.50	**0.50**	**-0.30**	0.5023	0.1118	0.00233	95.0	0.5022	0.1118	0.00223	95.0
	0.50	**0.50**	**-0.40**	0.5035	0.1118	0.00345	96.4	0.5035	0.1117	0.00354	96.4
	0.50	**0.50**	**-0.50**	0.5052	0.1117	0.00522	95.2	0.5054	0.1117	0.00536	95.0
500	**0.50**	**0.50**	**-0.01**	0.5029	0.1006	0.00286	96.2	0.5028	0.1006	0.00280	96.2
	0.50	**0.50**	**-0.10**	0.4991	0.1003	-0.00089	95.4	0.4988	0.1003	-0.00122	95.4
	0.50	**0.50**	**-0.20**	0.5010	0.1001	0.00103	95.8	0.5008	0.1001	0.00076	95.8
	0.50	**0.50**	**-0.30**	0.4994	0.1000	-0.00065	94.8	0.4993	0.1000	-0.00074	94.8
	0.50	**0.50**	**-0.40**	0.5031	0.0999	0.00308	94.2	0.5032	0.1000	0.00319	94.4

	0.50	**0.50**	**-0.50**	0.4998	0.0999	-0.00020	96.0	0.5000	0.1000	-0.00001	96.2

¥Standard Error; §Coverage Probability

β = Treatment effect on longitudinal outcome; γ = Treatment effect on survival outcome;

α = Association parameter between longitudinal and survival outcomes;

Note: The values of β (0.5) and γ (0.5) are fixed. However, the values of association parameter α are -0.01, -0.1, -0.2, -0.3, -0.4, and -0.5.

CHAPTER 6

DISCUSSION

6.1 Introduction

In this thesis, I applied random effects joint models (Rizopoulos, 2012; Wulfsohn and Tsiatis, 1997) to describe the longitudinal process and the survival process with competing risks failure time data. The longitudinal process was characterized by linear mixed (Laird and Ware, 1982) submodels, and the survival process was characterized by Cox proportional CSH (Cox, 1972) and proportional SDH (Fine and Gray, 1999) submodels. The dependency between the longitudinal and survival processes was considered using the underlying longitudinal value (Rizopoulos, 2010; Sweeting and Thompson 2011). The proportional SDH model (Fine and Gray, 1999) is a semi-parametric proportional regression model for the subdistribution hazard function (Li, Scheike, and Zhang, 2015). Thus, I used semi-parametric proportional CSH and SDH submodels for joint modeling. The EM algorithm was used to estimate unknown parameters of the model (Dempster et al., 1977; Rizopoulos, 2012; Wulfsohn and Tsiatis, 1997).

Most studies in joint modeling of longitudinal and survival data consider a single survival outcome and an assumption of independent censoring (Sweeting and Thompson 2011; Williamson et al., 2008). Some literature extends the methodology to allow for competing risks survival data (Deslandes and Chevret, 2010; Elashoff et al., 2007, 2008; Hu et al., 2009; Li et al., 2009; Williamson et al., 2008). However, most of them fitted only CSH submodels for competing risk events. Extension of the joint modeling approach to allow SDH submodel for competing risks has received limited attention to date. Only Deslandes and Chevret (2010)

considered SDH for competing risks submodels in their joint model. They applied the methodology in a study where the sequential organ failure assessment (SOFA) score and time of discharge and death for intensive care unit (ICU) patients were the longitudinal outcome and survival outcome, respectively.

6.2 Objective 1: To compare the two joint modeling approaches based on (i) Cox cause-specific hazards and (ii) subdistribution hazards via their application to real HIV/AIDS data.

Two time-to-event or survival outcomes cancer AIDS and non-cancer AIDS were defined in our study. Similar to Shiels et al. (2008, 2010), when cancer AIDS was the main event of interest, then non-cancer AIDS was the competing event and vice versa. Consequently, I considered Cox CSH and SDH competing risks submodels in the joint analyses. The first objective of this thesis was to compare results between the joint model with the CSH submodel and the joint model with the SDH submodel. For both cancer AIDS and non-cancer AIDS events, results were numerically different in the two joint models. For a cancer AIDS event, estimates of the intercept were 19.54 (se = 0.25) and 16.78 (se = 0.23) in the longitudinal submodels using CSH and SDH, respectively (Table 4.9). Age at diagnosis and CD4+ counts were negatively associated in the longitudinal submodel of a CSH-based joint model, while they were positively associated in the longitudinal submodel of an SDH-based joint model. In the survival submodels, point estimates and the corresponding 95% confidence intervals of the hazards ratio of cancer for MSM were 4.52 [95% CI: 1.51, 13.52] and 4.76 [95% CI: 1.54, 14.72] in the CSH-based joint model and the SDH-based joint model, respectively.

For a non-cancer AIDS event, estimates (standard errors) of MSM effects on CD4+ counts were 1.10 (se = 0.15) and 1.18 (se = 0.11) in the longitudinal submodels using CSH and SDH, respectively (Table 4.10). However, in the survival submodels, results were similar but not identical in CSH-based and SDH-based joint models.

From a methodological point of view, I conclude that results can be different in the CSH-based and SDH-based joint models. The magnitude of the differences in covariate parameter estimates could depend on the proportions of the competing risk events. The simulation study (Section 5.1) showed that the results could vary between CSH and SDH survival models depending on the proportions of the event of interest and the competing risk event. If we obtain different results in the CSH and SDH survival models, results can be different in their corresponding joint models as well. However, if results in CSH and SDH survival models are similar, the results between the CSH-based joint model and the SDH-based joint model may not be different. Therefore, in competing risks scenario, if results in the CSH and SDH survival models are different, I recommend joint modeling of longitudinal measurements using both CSH and SDH survival submodels. Then we can focus on the results in the model based on our objective or research question. For example, if our interest is to develop a clinical prediction model, we may consider results from the SDH-based joint model.

6.3 Objective 2: To examine the appropriateness of using cause-specific hazards and subdistribution hazards models based on the simulation study.

In my study, the results varied between the Cox CSH and SDH approaches for cancer AIDS. The results were similar but not identical in the two approaches for non-cancer AIDS.

Szychowski et al., 2010 reported that on some occasions, even though competing risk events are present but because of the fewer frequency, they have minimal impact on the SDH model parameter estimates. Berry, Ngo, Samelson, and Kiel (2010) indicated that if the percentage of individuals experiencing a competing risk is low, we may not get significantly different estimates in the CSH and SDH models. In my study, 22 (2.7%) individuals developed cancer-related AIDS, while 123 (15.0%) developed non-cancer AIDS. Thus, in the SDH model for non-cancer AIDS, the proportion of the competing event (cancer AIDS) was very low. This could be the reason similar results were obtained in the CSH and SDH models for non-cancer AIDS. I conducted a simulation study to verify the performance of the CSH and SDH techniques. For the simulation, I considered three scenarios using different cause-specific baseline hazard functions that allowed variation in the number event and the competing event in the simulated data. Proportions of cancer AIDS and non-cancer AIDS were higher in the simulated data compared to the real HIV data. The mean proportions of individuals with cancer AIDS, non-cancer AIDS were *14.6%, 20.6%,* respectively, in scenario 1; *15.8%, 25.0%,* respectively, in scenario 2; and *16.1%, 29.1%,* respectively, in scenario 3 (Table 5.2). I performed CSH and SDH analyses using the simulated data. For both events, the results in SDH models were substantially different from the CSH models in all three scenarios. Thus, results from the simulation study suggest that proportions of both the event and competing event influence the results in the CSH and SDH models. If the proportion of the competing event is not much lower than the proportion of the event of interest, the SDH model will provide different results than the CSH model.

The SDH model may not be necessary when the proportion of the competing event is very low, and the proportion of the main event of interest is relatively high. The threshold of the

minimum frequency of competing risk depends on the research question (Szychowski et al., 2010). However, we can evaluate this if we perform both the CSH and SDH analyses and compare the results (Szychowski et al., 2010). When analyzing the effects of clinical characteristics on competing risk events, Latouche et al. (2013) suggested using the Cox CSH and SDH models, presenting the results for all causes side by side. I also suggest routinely incorporating both CSH and SDH analyses in situations where competing risk events are common unless the proportions of the competing events are very low.

6.4 Objective 3. To investigate how magnitudes of association parameter between the longitudinal marker and time-to-event outcome influence parameter estimates obtained by conducting separate Cox regression analysis and linear mixed effects modeling from the simulation study.

In this objective, I want to determine if magnitudes of the association parameter between longitudinal and survival outcomes influence the parameter estimate in separate Cox regression models and the linear mixed model. In this simulation study, I considered both negative and positive association between longitudinal and survival outcomes. Sweeting and Thompson (2011) used 0.20 as the modest correlation between the longitudinal and survival processes in their simulation study. In my simulation study, I considered several magnitudes of positive and negative association between the two processes. I performed 500 replications, each simulation with n = 200, 400, and 500 individuals.

In the simulation study, when there was no association between longitudinal and survival outcomes, the mean estimate of the treatment effect on survival outcome in separate Cox

model was close to the true value (γ = 0.50), i.e. the bias of the estimate was very minimal. This indicates that separate Cox regression model provides an unbiased estimate when there is no association between the two outcomes. However, when there was an association, the mean estimate of the treatment effect in separate Cox model was different from the true value. In other words, separate Cox model provided bias estimate. When the association parameter, α, was set to 0.20, the bias of the estimate was 0.08 for all sample sizes. When the association parameter was 0.30, the bias of the estimate was 0.10. The bias of the parameter estimates increased as the magnitude of the association parameter increased, meaning that there was a trend. Hence, I conclude that when there is an association between longitudinal and survival outcomes, separate Cox regression analysis provides bias estimate. The magnitude of the bias depends on the magnitude of the association parameter.

When I used negative association between longitudinal and survival outcomes, the bias of the estimate in separate Cox models was larger. Thus, in a study (e.g. HIV/AIDS), if longitudinal and survival outcomes are negatively associated, we can make serious error statistical inference from separate Cox regression analysis. Please note that in HIV/AIDS studies, CD4+ count and time-to-AIDS or death are negatively associated. The early evolution of joint models was also based on HIV/AIDS clinical studies (DeGruttola and Tu, 1994; Faucett and Thomas, 1996; LaValley and DeGruttola, 1996; Pawitan and Self, 1993; Taylor et al., 1994; Tsiatis et al., 1995; Wulfsohn and Tsiatis, 1997).

In the separate linear mixed model, mean estimates of the treatment effect on longitudinal outcome are almost equal to the true value (β = 0.50) for all sample sizes and all association parameters. As I mentioned in Section 5.2.1, I simulated data based on the joint model where

the main interest is to estimate the treatment effect on the survival outcome while considering the impact of time-dependent repeatedly measured outcomes on the survival outcome (Ibrahim et al., 2010; Rizopoulos, 2012). Thus, the setting of the joint model in my simulation study could be the reason for obtaining almost unbiased estimates in separate mixed models.

6.5 Clinical significances:

In my study with the participants of lower CD4+ counts (≤ 500) at baseline, MSM had significantly higher hazards for cancer-related AIDS. Several studies (Gingues and Gill, 2006; Orem, Otieno, and Remick, 2004) reported that the incidence of Kaposi's Sarcoma (KS) and Non-Hodgkin's Lymphoma (NHL) has been significantly decreased, and survival from these cancers has been improved for most patients with the initiation of Highly Active Antiretroviral Therapy (HAART). However, KS was frequently identified in MSM in a study from Southern Alberta by Gingues and Gill (2006). Suárez-García et al. (2013) also observed the higher risk of KS among MSM in their recent study. MSM had higher hazards of NHL in another study reported by Bohlius et al. (2009). Thus, the results of my study were consistent with other study findings. In Canada, MSM represent about 50% of all the HIV-infected people (Challacombe, 2013). In my study, 481 (58.5%) participants were MSM. Therefore, my findings have consequences from a clinical perspective.

According to Gingues and Gill (2006), patients who do not come for HIV care after diagnosis and patients who do not receive antiretroviral therapy at the time of diagnosis are more likely to be diagnosed with AIDS-defining cancers. Clifford et al. (2005) observed that the use of HAART might lower the risk of KS and NHL. Bonnet et al. (2006) reported that patients with

higher HIV RNA levels for a long time have a higher risk of NHL. In my study, the median $\log_{10}$ HIV RNA level was significantly higher among MSM compared to the other group (4.8 vs. 4.4, respectively). Thus, higher HIV RNA levels could possibly make them more vulnerable to cancer-related AIDS. Since ARV treatment can keep the viral load at a low level, my results stress the importance of earlier ART initiation for the MSM risk group.

6.6 Strength and weakness

Strength:

Regarding data, my study data is from a large cohort consisting of three-quarters of HIV-infected people in Ontario (Rourke et al., 2013). This cohort is a representative sample of the HIV-infected population receiving HIV treatment in Ontario (Rourke et al., 2013). Follow-up time was longer for those with a satisfactory number of CD4+ counts.

Regarding methodology, I applied joint modeling methodology, which has become popular in biomedical research especially in HIV/AIDS studies. I studied this methodology in competing risks scenarios and applied both cause-specific hazards submodels and subdistribution hazards submodels. I performed simulation studies for competing risks models as well as for joint models.

Weakness:

In terms of data, although I used data from a large HIV cohort, participants who voluntarily participated in this study could be different from rest of the HIV-infected people in Ontario (Rourke et al., 2013). Hence the study may have recruitment bias (Rourke et al., 2013). Between the two time-to-event outcomes, the proportion of one outcome (cancer AIDS) was very low

(2.7%). Follow-up data for antiretroviral therapies, as well as detailed information on ARV adherence, were not available. Hence, I could not incorporate that information into the model.

I used the **JM** package from R (Rizopoulos, 2010) to fit joint models. However, all options are not available in the current package. For example, since I used the Cox PH model as our survival submodel in the joint analyses, I was not able to provide Martingale and Cox –Snell residual plots for survival submodel diagnostics in joint modeling. This area of joint model diagnosis remains as future research.

CHAPTER 7

CONCLUSION AND FUTURE RESEARCH

7.1 Conclusion

In this study, MSM had significantly higher risk than the other risk group for cancer AIDS (KS or NHL). However, for non-cancer AIDS, risks were not significantly different between MSM and the other risk group in the joint modeling analyses. Higher viral load at baseline could possibly make MSM vulnerable to cancer AIDS. Thus, my findings suggest earlier an ART initiation, as well as increased monitoring for the MSM HIV risk group, especially those with lower CD4+ counts at baseline (≤ 500). Regular HIV testing in individuals with HIV risk group is also vital, as untreated HIV infection represents a significant risk for cancer AIDS. Regular monitoring/screening for cancer is also important in individuals in the MSM HIV risk group.

In the competing risks scenario, depending on the proportion of the competing events, results could vary between the CSH and SDH survival models. Hence, if the results are different between CSH and SDH models, it is recommended to fit joint models with the CSH submodel as well as with the SDH submodel. We can then prefer results to the model according to our research question.

The bias in the estimate of the regression parameter in separate Cox regression analysis increases as the magnitude of the association between longitudinal and survival outcomes increases. Therefore, when longitudinal and survival outcomes are highly correlated, separate Cox regression analysis can provide serious error statistical inference about the regression parameters.

My main contributions in this thesis:

In this thesis, I applied joint modeling technique using subdistribution hazards in the HIV/AIDS study. I compared results between the two joint models using Cox cause-specific hazards and subdistribution hazards. From this study, I identified that MSM have a higher risk for cancer-related AIDS (KS or NHL) in spite of having higher CD4+ counts.

I simulated joint models using both positive and negative association between the longitudinal and survival outcomes considering a series of values for the association parameter. Through this study, I recommend decision strategies when a clinical study has competing risks in survival data with longitudinal repeated measurements.

7.2 Future research

In this thesis, I studied joint analysis using longitudinal continuous (Gaussian) outcomes for the longitudinal submodel. However, in some clinical studies, we may have longitudinal binary outcomes. For example, in the study of respiratory illness, respiratory status (good or poor) can be collected longitudinally, and asthma or dropout before asthma can be recorded as time-to-event outcomes. In the study of respiratory illness, dropout can be considered as a competing risk of asthma and, thus, joint modeling using both CSH and SDH survival submodels can be developed.

In this study, I considered single longitudinal outcomes (CD4+ counts) for the longitudinal submodel. The extension of this methodology is to use multiple longitudinal outcomes. In the study of patients with end-stage heart failure awaiting cardiac transplantation, bilirubin, and glomerular filtration rates can be collected longitudinally. Death before transplant and

transplant can be recorded as time-to-event or competing risks outcomes (Kim et al., 2006). In the HIV/AIDS study, in addition to CD4+ counts, viral load is also collected longitudinally (Lim et al., 2013). In the aforementioned studies, both longitudinal outcomes can be associated with time-to-event outcomes.

N. Li, Elashoff, G. Li, and Tseng (2012) studied joint analysis with multiple longitudinal responses and multiple failures time-to-event data (competing risks). However, they have used CSH frailty submodels for competing risks. Therefore, jointly modeling of multivariate longitudinal data and competing risks data using both CSH and SDH survival submodels is a future research topic of interest.

Most of the joint models in literature have concentrated on the development of models (Henderson et al., 2000; Rizopoulos et al., 2010; Wulfsohn and Tsiatis, 1997). Limited literature can be found in the area of joint model diagnostics (Dobson and Henderson, 2003; Rizopoulos et al., 2010). Rizopoulos et al. (2010) proposed residuals-plots based on multiple-imputation for joint models diagnostics. However, their method is applicable for parametric survival submodel. Still, there is a lot of scope for research in the area of joint model diagnostics.

In objective 2 of this thesis, I simulated competing risks data based on CSH models. I would like to simulate competing risks data based on SDH model in future.

REFERENCES

Ahmed, F. E., Vos, P. W., & Holbert, D. (2007). Modeling survival in colon cancer: a methodological review. *Molecular Cancer*, *6*, 15. http://doi.org/10.1186/1476-4598-6-15

AIDS.gov. (2010, November 16). Opportunistic Infections*. Opportunistic Infections and Their Relationship to HIV/AIDS*. Retrieved from https://www.aids.gov/hiv-aids-basics/staying-healthy-with-hiv-aids/potential-related-health-problems/opportunistic-infections/

AIDS.gov. (2015, August 27). *Stages of HIV infection*. Retrieved from https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/hiv-in-your-body/stages-of-hiv/

AIDS.gov. (2015, September 03). *Viral load*. Retrieved from https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/understand-your-test-results/viral-load/

AIDS.gov. (2015a, December 31). How do you get HIV or AIDS? *How is HIV spread?* Retrieved from https://www.aids.gov/hiv-aids-basics/hiv-aids-101/how-you-get-hiv-aids/

AIDS.gov. (2015b, December 31). *Symptoms of HIV*. Retrieved from https://www.aids.gov/hiv-aids-basics/hiv-aids-101/signs-and-symptoms/

AIDS.gov. (2016, July 14). CD4 count*. What is a CD4 count and why is it important*? Retrieved from https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/understand-your-test-results/cd4-count/

AIDSinfo. (2014, May 01). *Clinical guidelines portal*. Retrieved from https://aidsinfo.nih.gov/guidelines/html/1/adult-and-adolescent-arv-guidelines/458/plasma-hiv-1-rna--viral-load--and-cd4-count-monitoring

AIDSinfo. (2016, September 13). Education materials*. HIV overview*. Retrieved from https://aidsinfo.nih.gov/education-materials/fact-sheets/19/73/the-hiv-life-cycle

Aidsmap (n.d). *Zidovudine (AZT, Retrovir)*. Retrieved from http://www.aidsmap.com/resources/treatmentsdirectory/drugs/Zidovudine-AZT-iRetroviri/page/1730919/

Albert, P. S. and Shih, J. H. (2010). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics*, 66(3):983-987.

Allison, P. D. (2010). *Survival Analysis Using SAS: A Practical Guide, second edition*. Cary, NC/SAS Institute, Inc.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. Second International Symposium on Information Theory (eds. B. N. Petrov and F. Csaki), 267–281, Akademiai Kaido, Budapest.

Ancelle-Park, R. (1993). Expanded European AIDS case definition. *Lancet*, 341(8842):441.

Andersen, P. K., Abildstrom, S. Z., and Rosthoj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research*, 11(2):203-215.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study*. Annals of Statistics*, 10(4):1100-1120.

Asimow, L. A. and Maxwell, M. M. (2015). *Probability and statistics with applications: A problem solving text*. Winsted, CT/ACTEX Publications, Inc.

Austin, P. C., Lee, D. S., and Fine, J. P. (2016). Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation*, 133:601–609. DOI: 10.1161/CIRCULATIONAHA.115.017719.

Bakoyannis, G. and Touloumi, G. (n.d.). A Practical Guide on Modeling Competing Risk Data by Giorgos Bakoyannis and Giota Touloumi on behalf of CASCADE Collaboration. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.578.3091&rep=rep1&type=pdf

Bakoyannis, G. and Touloumi, G. (2010). Practical methods for competing risks data: A review. *Statistical Methods in Medical Research*, 21(3):257-272.

Balakrishnan, N. & Rao, C. R. (2004). *Handbook of Statistics: Advances in Survival Analysis*. Amsterdam, the Netherlands/Elsevier.

Bang, H., Chiu, Y.-L., Kaufman, J. S., Patel, M. D., Heiss, G., & Rose, K. M. (2013). Bias Correction Methods for Misclassified Covariates in the Cox Model: comparison offive correction methods by simulation and data analysis. *Journal of Statistical Theory and Practice*, *7*(2), 381–400. http://doi.org/10.1080/15598608.2013.772830

Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika*, 75(1):65-74.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software,* Volume 67, Issue 1. doi: 10.18637/jss.v067.i01

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine,* 24(11):1713-1723.

Berry, S. D., Ngo, L., Samelson, E. J., and Kiel, D. P. (2010). Competing risk of death: An important consideration in studies of older adults. *Journal of the American Geriatric Soc*iety, 58(4):783-787.

Bewick, V., Cheek, L., and Ball, J. (2004). Statistics review 12: Survival analysis. *Critical Care*, 8:389-394.

Beyersmann, J., Allignol, A., and Schumacher, M. (2012). *Competing Risks and Multistate Models with R.* London: Springer.

Beyersmann, J., Latouche, A., Buchholz, A., and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine*, 28(6):956-971.

Beyersmann, J. and Schumacher, M. (2007). Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in Medicine*, 26(7):1649-1651.

Bilmes, J. A. (1998, April). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Retrieved from http://melodi.ee.washington.edu/people/bilmes/mypapers/em.pdf

Bimali, M. and He, J. (2015). Association between Obesity and Cancer: An Analysis Using the Competing Risk Regression Approach. Advances in Epidemiology, Article ID 132961, 7 pages. doi:10.1155/2015/132961

Bohlius, J., Schmidlin, K., Costagliola, D., Fätkenheuer, G., May, M., Maria Caro-Murillo, A., … Egger, M. (2009). Incidence and risk factors of HIV-related non-Hodgkin's lymphoma in the era of combination antiretroviral therapy: a European multicohort study. *Antiviral Therapy*, 14(8): 1065-1074.

Bonnet, F., Balestre, E., Thiébaut, R., Morlat, P., Pellegrin, J. L., Neau, D., and Dabis, F. (2006). Factors Associated with the Occurrence of AIDS Related Non-Hodgkin Lymphoma in the Era of Highly Active Antiretroviral Therapy: Aquitaine Cohort, France. Clinical Infectious Diseases, 42:411-7.

Borman, S. (2006). *The Expectation Maximization Algorithm. A short tutorial.* Retrieved from https://www.cs.utah.edu/~piyush/teaching/EM_algorithm.pdf

Breastcacer.Org (2017, January 31). *Where Can Breast Cancer Come Back or Metastasize*? Retrieved from http://www.breastcancer.org/symptoms/types/recur_metast/where_recur

Brombin, C., Serio, C. D., and Rancoita, P. MV. (2014). Joint modeling of HIV data in multicenter observational studies: A comparison among different approaches. *Statistical Methods in Medical Research*, 0(0) 1-16. DOI: 10.1177/0962280214526192

Brown, E. R. and Ibrahim, J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, 59(2):221-228.

Brown, E. R., Ibrahim, J. G., and DeGruttola, V. (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64-73.

Brown, H. and Prescott, R. (2006). *Applied mixed models in medicine* (2nd ed.). West Sussex, England: John Wiley & Sons, Ltd.

Carter, M. (2016a, March). Aidsmap. *Factsheet CD4 cell counts*. Retrieved from http://www.aidsmap.com/CD4-cell-counts/page/1044596/

Carter, M. (2016b, March). Aidsmap. *Factsheet primary infection*. Retrieved from http://www.aidsmap.com/Primary-infection/page/1044761/

Castro, H. G., Ward, J. W., Slutska, L., Baehler, J. W., Jaffo, H. W., and Bertrelman, R. L. (1992). 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *MMWR Recommendations and Reports*, 41: 1-19.

CATIE (n.d.). *A practical guide to HIV Drug Treatment for People Living with HIV*. Retrieved from http://www.catie.ca/en/practical-guides/hiv-drug-treatment/2-hiv-and-aids-basics/2-2

CDC. (2008, November 20). Appendix A. AIDS-Defining Conditions. *MMWR Recommendations and Reports*, 57: 9. Retrieved from http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5710a2.htm

CDC. (2016, August 25). HIV/AIDS. *Opportunistic infections*. Retrieved from http://www.cdc.gov/hiv/basics/livingwithhiv/opportunisticinfections.html

Challacombe, L. (2013). *The epidemiology of HIV in Canada*. Retrieved from http://www.catie.ca/sites/default/files/epi-hiv-can-en.pdf

Charurat, M., Oyegunle, M., Benjamin, R., Habib, A., Eze, E., Ele, P., Ibanga, I., … Blattner, W. (2010). Patient retention and adherence to antiretrovirals in a large antiretroviral therapy program in Nigeria: a longitudinal analysis for risk factors. *PLoS One,* 5(5): e10584.

Cherney, K. (2014, September 15). *A Timeline of HIV Symptoms*. Retrieved from http://www.healthline.com/health/hiv-aids/hiv-symptoms-timeline#Overview1

Chernick, M. R. and Friis R. H. (2003). *Introductory Biostatistics for the health sciences. Modern application including bootstrap*. Hoboken, New Jersey: John Wiley & Sons.

Chi, Y. Y. and Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62(2):432-445.

Clifford, G. M., Polesel, J., Rickenbach, M., Dal Maso, L., Keiser, O., Kofler A., Rapiti, E.,… Franceschi, S. (2005). Cancer risk in the Swiss HIV Cohort Study: associations with immunodeficiency, smoking, and highly active antiretroviral therapy. *Journal of the National Cancer Institute*, 97(6):425-32.

Collett, D. (2003). *Modelling Survival Data in Medical Research* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Comiskey, C., Dempsey, O., Simic, D., and Baroš, S. (2013). Injecting drug users, sex workers and men who have sex with men: a national cross-sectional study to develop a framework and prevalence estimates for national HIV/AIDS programmes in the Republic of Serbia. *BMJ Open*, 3:e002203. doi:10.1136/bmjopen-2012-002203

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, Series B, 34(2):187-220.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society* B, 30(2):248-275.

Dafni, U. G. (1993). *Evaluating Surrogate markers of clinical outcome when measured with error.* (Unpublished doctoral dissertation). Harvard School of Public Health, Cambridge, MA.

Dafni, U. G. and Tsiatis, A. A. (1998). Evaluating surrogate markers of clinical outcome measured with error. *Biometrics,* 54(4):1445-1462.

DeGruttola, V. and Tu, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50(4):1003–1014.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* Series B, 39(1):1-38.

Denning, P. and DiNenno, E. (2015, June 23). *Communities in Crisis: Is There a Generalized HIV Epidemic in Impoverished Urban Areas of the United States*? Retrieved from http://www.cdc.gov/hiv/group/poverty.html

Deslandes, E. and Chevret, S. (2010). Joint modeling of multivariate longitudinal data and the dropout process in a competing risk setting: application to ICU data. *BMC Medical Research Methodology*, 10:69.

Diaz, L. C. (2014). *Joint Modelling for Longitudinal and Time-to-Event Data*. Application to Liver Transplantation Data. Master's thesis. University of Santiago de Compostela.

Diggle, P., Heagetry P., Liang, K.Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press. Second eidition.

Dignam, J. J., Zhang, Q., & Kocherginsky, M. N. (2012). The Use and Interpretion of Competing Risks Regression Models. *Clinical Cancer Research*, *18*(8), 2301–2308. http://doi.org/10.1158/1078-0432.CCR-11-2097

Ding, J. and Wang, J. L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, 64(2):546-556.

Dobson, A. and Henderson, R. (2003). Diagnostics for joint longitudinal and dropout time modeling. *Biometrics,* 59(4):741-751.

Duchateau, L., Janssen, P., and Rowlands, J. (1998). *Linear mixed models. An introduction with applications in veterinary research*. ILRI (International Livestock Research Institute), Nairobi, Kenya. 159 pp.

Dupuy, J. F. and Mesbah, M. (2002). Joint modeling of event time and nonignorable missing longitudinal data. *Lifetime data analysis*, 8, 99-115.

Ebrahim, S. H., Abdullah, A. S., McKenna, M., and Hamers, F. F. (2004). AIDS defining cancers in Western Europe, 1994–2001. *AIDS Patient Care STDS*, 18(9): 501-508.

Edurant (2016, August 25). *What is HIV?* Retrieved from http://www.edurant.com/patients/hiv-background-treatment/what-is-hiv

Elashoff, R. M., Li, G., and Li, N. (2007). An approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in Med*icine, 26(14):2813-2835.

Elashoff, R. M., Li, G., and Li, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64(3):762-771.

Elder, C. (2013, November 18). Treatments and Cures for HIV/AIDS. Retrieved from https://prezi.com/3h5gcmodxvzi/treatments-and-cures-for-hivaids/

Faucett, C. L., Schenker, N., and Taylor, J. M. G. (2002). Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics*, 58(1):37-47.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modeling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, 15(15):1663-1685.

Fine, J. P. (2001). Regression modeling of competing crude failure probabilities. *Biostatistics*, 2(1):85-97.

Fine, J. P., and Gray, R. J. (1999). A proportional hazards model for subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496-509.

Finucane, M. M., Samet, J. H., and Horton, N. J. (2007). Translational methods in biostatistics: linear mixed effect regression models of alcohol consumption and HIV disease progression over time. *Epidemiologic Perspectives & Innovations*, 4:8. doi:10.1186/1742-5573-4-8.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis* (2nd ed.). New York: Wiley.

Fitzmaurice , G. M. and Ravichandran, C. (2008). A Primer in Longitudinal Data Analysis. *Circulation*,118:2005-2010. DOI: 10.1161/CIRCULATIONAHA.107.714618

Fox, J. (2008). *Cox Proportional-Hazards Regression for Survival Data. Appendix to An R and S-PLUS Companion to Applied Regression*. Retrieved from https://socserv.socsci.mcmaster.ca/jfox/Books/Companion-1E/appendix-cox-regression.pdf

Fox, J. and Weisberg, S. (2011). *Cox Proportional-Hazards Regression for Survival Data in R.* An Appendix to An R Companion to Applied Regression, Second Edition. Retrieved from https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Cox-Regression.pdf

Gaynor, J. J., Feuer, E. J., Tan, C. C., Wu, D. H., Little, C. R., Straus, D. J., Clarkson, B. D., and Brennan, M. F. (1993). On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association*, 88(422): 400-409.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Geskus, R. B. (2016). *Data Analysis with Competing Risks and Intermediate States*. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.

Gingues, S and Gill, M. J. (2006). The impact of highly active antiretroviral therapy on the incidence and outcomes of AIDS-defining cancers in Southern Alberta. *HIV Medicine*, 7(6):369-377.

Givens, G. H. & Hoeting, J. A. (2013). *Computational statistics. Second edition*. Hoboken, NJ: John Wiley & Sons, Inc.

Gooley, T. A., Leisenring, W., Crowley, J., and Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine,* 18(6):695-706.

Grambauer, N., Schumacher, M., and Beyersmann, J. (2010). Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Statistics in Medicine*, 29(7-8):875-884.

Gray, R. (1988). A class of K-Sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics*, 16(3):1141-1154.

Green, E. C., and Ruark, A. H. (2016). *AIDS*, *Behavior, and Culture. Understanding evidence-based prevention*. New York, NY: Routledge/Taylor and Francis group.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58(1):16-24.

Haller, B., Schmidt, G. & Ulm, K. (2012). Applying competing risks regression models: an overview*. Lifetime Data Anal*. DOI 10.1007/s10985-012-9230-8

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320-338.

Healthline (2015, December 22). *Acute HIV infection*. Retrieved from http://www.healthline.com/health/acute-hiv-infection#Overview1

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465-480.

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16:117. DOI: 10.1186/s12874-016-0212-5

Hinchliffe, S. R. and Lambert, P. C. (2013). Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC Medical Research Methodology*, 13:13. doi:10.1186/1471-2288-13-13

HIV Monitoring. (n.d.). *About HIV*. Retrieved from http://www.hiv-monitoring.nl/english/patients-and-public/about-hiv/

Hogg, R. S., Yip, B., Kully, C., Craib, K. J., O'Shaughnessy, M. V., Schechter, M. T., and Montaner, J. S. (1999). Improved survival among HIV-infected patients after initiation of triple-drug antiretroviral regimens. *Canadian Medical Association Journal*, 160(5):659-665.

Holt, J. D. (1978). ‘‘Competing risk analyses with special reference to matched pair experiments,’’ Biometrika vol. 65 pp. 159-165.

Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, 62(4):1037-1043.

Hu, W. H., Li, G., and Li, N. (2009). A Bayesian approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in Med*icine, 28(11):1601-1619.

Huang, Y., Dagne, G., and Wu, L. (2011). Bayesian inference on joint models of HIV dynamics for time-to event and longitudinal data with skewness and covariate measurement errors. *Statistics in Medicine*, 30(24):2930-2946.

Hunt, J. R. and White, E. (1988). Retaining and Tracking Cohort Study Members. *Epidemiologic Reviews*, Vol. 20, No. 1.

Ibrahim, J. G., Chen, M., and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica*, 14: 863-883.

Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796-2801.

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test (Madrid, Spain)*, *18*(1), 1–43. http://doi.org/10.1007/s11749-009-0138-x

Kagan, J. M., Sanchez, A. M., Landay, A., & Denny, T. N. (2015). A Brief Chronicle of CD4 as a Biomarker for HIV/AIDS: A Tribute to the Memory of John L. Fahey. *Forum on Immunopathological Diseases and Therapeutics*, *6*(1-2), 55–64. http://doi.org/10.1615/ForumImmunDisTher.2016014169.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data* (2nd ed.). Wiley: New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457-481.

Kim, H. T. (2007). Cumulative incidence in competing risks data and competing risks regression analysis. *Clinical Cancer Research*, 13(2.1):559-565.

Kim, W. R., Therneau, T. M., Benson, J. T., Kremers, W. K., Rosen, C. B., Gores, G. J., and Dickson, E. R. (2006). Deaths on the Liver Transplant Waiting List: An Analysis of Competing Risks. *Hepatology*, volume 43, issue 2.

Kleinbaum, D. G. and Klein, M. (2012). Survival analysis. A self-learning text. Third edition. Springer.

Klein, J. P. and Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudo values of the cumulative incidence function. *Biometrics*, 61(1):223-229.

Knott, L. (2015, November 04). *HIV and AIDS*. Retrieve from http://patient.info/health/hiv-and-aids

Koller, M. T., Raatz, H., Steyerberg, E. W., and Wolbers, M. (2011). Competing risks and the clinical community: irrelevance or ignorance? *Stat Med*., 31:1089-1097. doi: 10.1002/sim.4384.

Kuk, D. & Varadhan, R. (2013). Model selection in competing risks regression. *Statist. Med,* 32 3077-3088. DOI: 10.1002/sim.5762

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963-974.

Lange, K. (2010). *Numerical analysis for statisticians. Second edition.* New York, NY: Springer.

Langford, S., Ananworanich, J., and Cooper, D. (2007). Predictors of disease progression in HIV infection: a review. *AIDS Research and Therapy*, 4:11. doi:10.1186/1742-6405-4-11.

Latouche, A., Allignol, A., Beyersmann, J., Labopind, M., and Fine, J. P. (2013). A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology*, 66:648-653.

Latouche, A., Boisson, V., Chevret, S. and Porcher, R. (2007), Misspecified regression model for the subdistribution hazard of a competing risk. Statist. Med., 26: 965–974. doi:10.1002/sim.2600

Lau, B., Cole, S. R., and Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *Am J Epidemiol*., 170:244-256. doi: 10.1093/aje/kwp107

Lavalley, M. P. and DeGruttola, V. (1996). Model for empirical Bayes estimators of longitudinal CD4 counts. *Statistics in Medicine,* 15(21-22):2289-2305.

Lewis, T. H. (2017). *Complex survey data analysis with SAS*^{R}. Boca Raton, FL: Chapman and Hall/CRC Press.

Li, J. and Ma, S. (2013). *Survival analysis in medicine and genetics*. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.

Li, N., Elashoff, R. M., and Li, G. (2009). Robust joint modeling of longitudinal measurements and competing risks failure type data. *Biometrical Journal*, 51(1):19-30.

Li, N., Elashoff, R. M., Li, G., and Tseng, C. H. (2012). Joint analysis of bivariate longitudinal ordinal outcomes and competing risks survival times with nonparametric distributions for random effects. *Stat Med*, 31(16):1707-21.

Li, J., Scheike, T. H., & Zhang, M.-J. (2015). Checking Fine and Gray Subdistribution Hazards Model with Cumulative Sums of Residuals. *Lifetime Data Analysis*, *21*(2), 197–217. http://doi.org/10.1007/s10985-014-9313-9

Liang, K. -Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika,* 73(1):13-22.

Lim, H. J., Mondal, P., and Skinner, S. (2013). Joint modeling of longitudinal and event time data: application to HIV study. *Journal of Medical Statistics and Informatics,* 1:1.

Littell, R. C., Pendergast, J., and Natarajan, R. (2000). Tutorial in Biostatistics. Modelling covariance structure in the analysis of repeated measures data. *Statist. Med*., 19:1793-1819.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Little, R. J. A. and Rubin, D. (2002). Statistical Analysis with Missing Data. 2nd edition. John Wiley & Sons, New York.

Lunn, M. and McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics*, 51(2):524-532.

Maartens, G., Celum, C., and Lewin, S. R. (2014). HIV infection: epidemiology, pathogenesis, treatment, and prevention. *Lancet*, 384(9939): 258-71.

McCrink, L., Marshall, A. H., & Cairns, K. (2011). Joint Modelling of Longitudinal and Survival Data: A Comparison of Joint and Independent Models. *Int. Statistical Inst*.: Proc. 58th World Statistical Congress, 2011, Dublin (Session CPS044)

McMurchy, D., Challacombe, L., Edmiston, M., … Rachlis, A. (2010). Gaining trust, ensuring security: the evolution of an Ontario HIV cohort study. In C. M. Flood (Ed.), *Data Data Everywhere: Access and Accountability?* Montreal and Kingston: Montréal/Kingston: McGill-Queen's University Press.

MedicineNet.com (2016, May 13). *Definition of opportunistic infection*. Retrieved from http://www.medicinenet.com/script/main/art.asp?articlekey=11772

Metropolis, N. and Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association*, 44.

Mishra, S., Sgaier, S. K., Thompson, L. H., Moses, S., Ramesh, B. M., Alary, M., Wilson, D., and Blanchard, J. F. (2012). HIV Epidemic Appraisals for Assisting in the Design of Effective Prevention Programmes: Shifting the Paradigm Back to Basics. PLoS ONE 7(3): e32324. doi:10.1371/journal.pone.0032324

Murawska, M. (2013). *Extensions in Joint Modeling of Survival and Longitudinal Outcomes* (Doctoral dissertation). Erasmus Medical Center, Rotterdam, the Netherlands.

Nall, R. (2016). *How HIV affects the body*. Retrieved from http://www.healthline.com/health/hiv-aids/how-hiv-affects-the-body#Overview1

Orem, J., Otieno, M. W., and Remick, S. C. (2004). AIDS-associated cancer in developing nations. *Current Opinion in Oncology*, 16(5):468-476.

Pauler, D. K. and Finkelstein, D. M. (2002). Predicting time to prostate cancer recurrence based on joint models for non-linear longitudinal biomarkers and event time outcomes. *Statistics in Medicine*, 21:3897-3911.

Pawitan, Y. and Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of the American Statistical Association,* 88(423):719-726.

Pepe, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association,* 86(415)**:**770-778.

Persson, I. (2002). *Essays on the Assumption of Proportional Hazards in Cox Regression*. Doctoral dissertation. Uppsala University, Uppsala, Sweden.

PHAC. (2014, November 28). *Chapter 1: National HIV Prevalence and Incidence Estimates for 2011*. Retrieved from http://www.phac-aspc.gc.ca/aids-sida/publication/epi/2010/1-eng.php

Philipson, P., Sousa, I., Diggle, P. (2012, March 29). *joineR: Joint modelling of repeated measurements and time-to-event data*. Retrieved from http://nrl.northumbria.ac.uk/6945/4/joineR_vignette.pdf

Pintilie, M. (2006). *Competing Risks: A Practical Perspective.* New York: Wiley.

Pintilie, M. (2007). Analysing and interpreting competing risk data. *Statistics in Medicine,* 26:1360-1367.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics,* 34(4):541–554.

Proust-Lima, C., & Taylor, J. M. G. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics (Oxford, England)*, *10*(3), 535–549. http://doi.org/10.1093/biostatistics/kxp009

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389-2430.

Raboud, J. M., Abdurrahman, Z. B., Major, C., Millson, P., Robinson, G., Rachlis, A., and Bayoumi, A. M. (2005). Nonfinancial factors associated with decreased viral load testing in Ontario, Canada. *Journal of Acquirred Immune Deficiency Syndromes*, 39(3):327–332.

Rizopoulos, D. (2010). JM: an R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1-33.

Rizopoulos, D. (2011). Dynamic predictions and properspective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819-829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data With Applications in R.* Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society,* Series B, 71(3):637-654.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*, 66(1):20-29.

Rourke, S. B., Gardner, S., Burchell, A. N., Raboud, J., Rueda, S., Bayoumi, A. M., Loutfy, M., … Rachlis, A. (2013). Cohort Profile: The Ontario HIV Treatment Network Cohort Study (OCS). *International Journal of Epidemiology*, 42(2):402-411.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics,* 43(3):429–467.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Salkind, N. J. and Rasmussen, K. (2008). *Encyclopedia of Educational Psychology, Volume 1*. Thousand oaks, California: Sage Publications, Inc.

SAS (n.d.). *SAS/STAT software. Longitudinal Data Analysis*. Retrieved from https://support.sas.com/rnd/app/stat/procedures/LongitudinalAnalysis.html

Satagopan, J. M., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D., & Auerbach, A. D. (2004). A note on competing risks in survival data analysis. *British Journal of Cancer*, *91*(7), 1229-1235. http://doi.org/10.1038/sj.bjc.6602102

Sattar, A., Sinha, S.K., Argyropoulos, C., & Unruh, M. (2012). Joint Modeling of All-Cause Mortality and Longitudinally Measured Serum Albumin. Progress in Applied Mathematics, 4 (2):182-195.

Saville, B. R., Herring, A. H., and Koch, G. G. (2009). A robust method for comparing two treatments in a confirmatory clinical trial via multivariate time-to-event methods that jointly incorporate information from longitudinal and time-to-event data. *Statistics in Medicine*, 29(1):75-85.

Scrucca, L., Santucci, A. & Aversa, F. (2010). Regression modeling of competing risk using R: an in depth guide for clinicians. *Bone Marrow Transplantation*, 45, 1388-1395. doi:10.1038/bmt.2009.359

Self, S. and Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. In N.P. Jewell, K. Dietz, and V.T. Farewell (Eds.), *AIDS Epidemiology: Methodological Issues*. Boston: BirkhÄauser.

SFAF. (n.d.). *Is there a cure for HIV or AIDS*? Retrieved from http://sfaf.org/hiv-info/basics/is-there-a-cure-for-hiv-aids.html

Shahapur, P. R., & Bidri, R. C. (2014). Recent trends in the spectrum of opportunistic infections in human immunodeficiency virus infected individuals on antiretroviral therapy in South India. *Journal of Natural Science, Biology, and Medicine*, *5*(2), 392–396. http://doi.org/10.4103/0976-9668.136200

Shiels, M. S., Cole, S. R., Chmiel, J. S., Margolick, J., Martinson, J., Zhang, Z. F., and Jacobson, L. P. (2010). A comparison of *ad hoc* methods to account for non-cancer AIDS and deaths as competing risks when estimating the effect of HAART on incident cancer AIDS among HIV-infected men. *Journal of Clinical Epidemiology*, 63(4):459-467.

Shiels, M. S., Cole, S. R., Wegner, S., Armenian, H., Chmiel, J. S., Ganesan, A., Marconi, V. C., … Crum-Cianflone, N. F. (2008). Effect of HAART on incident cancer and noncancer AIDS events among male HIV seroconverters. *Jounal of Acquired Immune Deficiency Syndromes*, 48(4):485-90.

Shiels, M. S., Pfeiffer, R. M., Gail, M. H., Hall, H. I., Li, J., Chaturvedi, A. K., Bhatia, K., Uldrick, T. S., Yarchoan, R., Goedert, J. J., Engels, E. A. (2011). Cancer Burden in the HIV-Infected Population in the United States. J Natl Cancer Inst, 103:753-762. DOI: 10.1093/jnci/djr076

Singh, R., & Mukhopadhyay, K. (2011). Survival analysis in clinical trials: Basics and must know areas. *Perspectives in Clinical Research*, *2*(4), 145–148. http://doi.org/10.4103/2229-3485.86872

Song, H. (2013*). Joint Modelling of Longitudinal Quality of Life Measurements and Survival Data in Cancer Clinical Trials*. Doctoral dissertation. Queen's University, Kingston, Canada.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics,* 58(4):742-753.

Spano, J.-P., Costagliola, D., Katlama, C., Mounier, N., Oksenhendler, E., and David Khayat, D. (2008). AIDS-related malignancies: State of the art and therapeutic challenges. *Journal of Clinical Oncology*, 26(29):4834-4842.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003). *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge. URL http://www.mrc-bsu.cam.ac.uk/bugs/.

Suárez-García, I., Jarrín, I., Iribarren, J. A., López-Cortésd, L. F., Lacruz-Rodrigoe, J., Masiáf, M., Gómez-Sirventg, J. L., … Amob J. D. (2013). Incidence and risk factors of AIDS-defining cancers in a cohort of HIV-positive adults: Importance of the definition of incident cases. *Enfermedades Infecciosas y Microbiología Cliníca*, 31(5):304-312.

Sullivan L. M. (2012). *Essentials of Biostatistics in public health*. Sudbury, MA/Jones and Bartlett learning.

Sweeting, M. J. and Thompson, S. G. (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53(5):750-763.

Szychowski, J. M., Roth, D. L., Clay, O. J., & Mittelman, M. S. (2010). Patient Death as a Censoring Event or Competing Risk Event in Models of Nursing Home Placement. *Statistics in Medicine*, *29*(3), 371–381. http://doi.org/10.1002/sim.3797

Taylor, J. M. G., Cumberland, W. G., and Sy, J. P. (1994). A stochastic model for analysis of longitudinal data. *Journal of the American Statistical Association,* 89(427):727-736.

Terrera, G. M., Piccinin, A. M., Johansson, B., Matthews, F., & Hofer, S. M. (2011). Joint Modeling of Longitudinal Change and Survival: An Investigation of the Association Between Change in Memory Scores and Death. *GeroPsych*, *24*(4), 177–185. http://doi.org/10.1024/1662-9647/a000047

Thaczuk, D. (n.d). *Managing your health: a guide for people living with HIV*. Retrieved from http://www.catie.ca/en/practical-guides/managing-your-health/2

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1):147-160.

Touloumi, G., Pantazis, N., Babiker, A. G., Walker, S. A., Katsarou, O., Karafoulidou, A., Hatzakis, A., Porter, K. (2004). Differences in HIV RNA levels before the initiation of antiretroviral therapy among 1864 individuals with known HIV-1 seroconversion dates. AIDS. 18(12):1697-705.

Tsiatis, A. A. (1999*). Competing Risks. Encyclopedia of Biostatistics*. New York: Wiley.

Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14:809-834.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association,* 90(429): 27-37.

UNAIDS (1998, October). *HIV-related opportunistic diseases*. Retrieved from http://www.who.int/hiv/pub/amds/opportu_en.pdf

UNAIDS. (2011). *Securing the future today.* Retrieved from http://www.unaids.org/sites/default/files/media_asset/20110727_JC2112_Synthesis_report_en_0.pdf

UNAIDS. (2013). *AIDS by the numbers*. Retrieved from http://www.unaids.org/en/resources/campaigns/globalreport2013/factsheet

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Verbeke, G., Molenberghs, G., and Beunckens, C. (2008). Formal and Informal model selection with incomplete data. *Statistical Science*, 23(2):201-218.

Volberding, P. A., Greene, W. C., Lange, J. M. A., Gallant, J. E., and Sewankambo, N. (Ed.). (2012). *SANDE's HIV/AIDS Medicine: Medical Management of AIDS* (2nd ed.). China: Elsevier/Saunders.

Vyas, J. M. (2015, May 01). MedlinePlus. *HIV/AIDS*. Retrieved from https://www.nlm.nih.gov/medlineplus/ency/article/000594.htm

Wang, W. (2004). Proportional hazards regression models with unknown link function and time-dependent covariates. *Statistica Sinica*, **14**, 885-905.

Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, 96(455):895-905.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician*, volume 39, issue 2, 95-101.

WebMD (n.d.). *CD4+ count*. Retrieved from http://www.webmd.com/hiv-aids/cd4-count#1

Weibull, W. (1951). A statistical distribution function of wide applicability*. J. Appl. Mech,* 18 (3): 293-297*.*

WHO (n.d.). *Antiretroviral therapy*. Retrieved from http://www.who.int/topics/antiretroviral_therapy/en/

WHO (2012, July 18). *'Strategic use' of HIV medicines could help end transmission of virus*. Retrieved from http://www.who.int/mediacentre/news/releases/2012/hiv_medication_20120718/en/

WHO (2016, November). *HIV/AIDS*. Retrieved from http://www.who.int/features/qa/71/en/

Williamson, P. R., Kolamunnage-Dona, R., Philipson, P., and Marson, A. G. (2008). Joint modeling of longitudinal and competing risks data. *Statistics in Medicine*, 27(30):6426-6438.

Wu, L., Liu, W., and Hu, X. J. (2010). Joint inference on HIV viral dynamics and immune suppression in presence of measurement errors. *Biometrics*, 66(2):327-335.

Wu, L., Liu, W., Yi, G. Y., and Huang, Y. (2012). Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues. *Journal of Probability and Statistics*. doi: http://dx.doi.org/10.1155/2012/640153.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error*. Biometrics*, 53(91):330-339.

Xu, J. and Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society,* Series C, 50(3):375-387.

Yao, F. (2008). Functional approach of flexibly modelling generalized longitudinal data and survival time. *Journal of Statistical Planning and Inference*, 138: 995-1009.

Ye, W., Lin, X., and Taylor, J. M. G. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data—a two-stage regression calibration approach. *Biometrics*, 64(4):1238-1246.

Yu, M., Law, N. J., Taylor, J. M. G., and Sandler, H. M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, 14:835-862.

Zeng, D. and Cai, J. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Annals of Statistics*, 33(5):2132-2163.

Zhang, M.-J., Zhang, X., & Scheike, T. H. (2008). Modeling cumulative incidence function for competing risks data. *Expert Review of Clinical Pharmacology*, *1*(3), 391–400. doi:10.1586/17512433.1.3.391